Cena 15,43 zł
(VAT 8%)

Indeks 371262
ISSN 0033-2372

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

PRZEGLĄD STATYSTYCZNY

STATISTICAL REVIEW

TOM 66

4

2019

GUS

Information for Authors

Przegląd Statystyczny (*Statistical Review*) publishes original research papers on theoretical and empirical topics in statistics, econometrics, mathematical economics, operational research, decision sciences and data analysis. The manuscripts considered for publication should significantly contribute to the theoretical aspects of the aforementioned fields or shed new light on the practical applications of these aspects. Manuscripts reporting important results of research projects are particularly welcome. Review papers, shorter papers reporting on major conferences in the field and reviews of seminal monographs are eligible for submission but only on the Editor's request.

Since 1 May 2019, the journal has been publishing articles in English.
Any spelling style is acceptable as long as it is consistent within the manuscript.

All work should be submitted to the journal through the ICI Publishers Panel (https://editors.publisherspanel.com/pl.ici.ppanel-app-war/ppanel/index).

For details of the submission process and editorial requirements please visit https://ps.stat.gov.pl/ForAuthors.

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

PRZEGLĄD STATYSTYCZNY

STATISTICAL REVIEW

TOM 66

4

2019

WARSZAWA 2019

ADRES REDAKCJI / EDITORIAL OFFICE
Uniwersytet Gdański, ul. Armii Krajowej 101, 81-824 Sopot

Redakcja językowa / Language editing: Wydział Czasopism Naukowych, Główny Urząd Statystyczny

Strona internetowa /Website: ps.stat.gov.pl

ISSN 0033-2372

e-ISSN 2657-9545

Indeks 371262

Zakład Wydawnictw Statystycznych

ZAKŁAD WYDAWNICTW STATYSTYCZNYCH
al. Niepodległości 208, 00-925 Warszawa, tel. 22 608 31 45
Andrzej Paluchowski (redaktor techniczny), Marzena Jędrzejewska (skład i łamanie)

Informacje w sprawie sprzedaży czasopisma tel.: 22 608 32 10, 22 608 38 10

Zam. 152/2019 — nakład 220 egz.

SPIS TREŚCI

SPRAWOZDANIA

CONTENTS

REPORTS

DOI: 10.5604/01.3001.0014.0951

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 4 – 2019

Magdalena BARSKA[1]

Analiza popytu w przemyśle hutniczym – zastosowanie modelu ze zmiennymi ukrytymi[2]

Streszczenie. *Na popyt w przemyśle hutniczym wpływa wiele czynników. Nie wszystkie można zidentyfikować i zmierzyć. W artykule przedstawiono wyniki analizy popytu dla wybranego przedsiębiorstwa w latach 2010–2014. Celem przedstawionego badania jest budowa ukrytego modelu łańcuchów Markowa, który odzwierciedli punkty zwrotne zapotrzebowania na wyroby hutnicze oraz umożliwi prognozę wielkości tego zapotrzebowania. Zbadano własności prognostyczne i stabilność modelu. Przeprowadzono symulację polegającą na wygenerowaniu dużej liczby szeregów dla zadanych parametrów modelu i sprawdzeniu ich własności. Najlepiej dopasowanym modelem okazał się trójstanowy model ukrytych łańcuchów Markowa. Za jego pomocą opisano stany potencjalnie kształtujące wielkość popytu. Uwzględnienie stanu przejściowego pozwoliło uchwycić sygnał nadchodzącej zmiany pomiędzy fazami wzrostu i spadku. Zaproponowany model ukrytych łańcuchów Markowa drugiego rzędu może być alternatywą dla tradycyjnych metod analizy szeregów czasowych. Wyznaczona prognoza informuje o kształtowaniu się trendu i stanowi wskazówkę co do punktów zwrotnych koniunktury.*

Słowa kluczowe: prognozowanie, popyt, zmienne ukryte, modele ukrytych łańcuchów Markowa

Analysis of demand in steel and iron industry – latent variables model

Abstract. *Demand in the steel and iron industry is influenced by multiple factors. Not all of them can be identified and measured. The paper presents the results of the analysis of the levels of demand achieved by a selected enterprise from this sector in the years 2010–2014. The aim of the study is to build a hidden Markov model which would reflect the turning points of this demand, thus making it possible to forecast its future levels. The model's forecasting properties and stability have been examined. A simulation has been carried out that involved generating a high number of series for selected model parameters and checking their properties. This demonstrated that a three-state second order hidden Markov model was most relevant to the purpose of the study. Thanks to the model's application, it was possible to describe states which could potentially shape the demand. Moreover, taking the transition state into consideration allowed spotting the signal about the upcoming replacement of the growth phase with the decline phase, and vice versa. The presented second order hidden Markov model can serve as an alternative to the traditional methods of the analysis of time series. The forecast generated by the model informs about the shaping of a trend in demand and serves as an indication of the shifts in economic cycles.*

Keywords: forecasting, demand, latent variables, hidden Markov models

JEL Classification: C15, C51

[1] Szkoła Główna Handlowa w Warszawie, Kolegium Analiz Ekonomicznych, al. Niepodległości 162, 02-554 Warszawa, e-mail: d09a1997@doktorant.sgh.waw.pl, ORCID: https://orcid.org/0000-0002-6410-7929.

[2] Badanie opisane w artykule zostało zrealizowane w ramach Badań Młodych Naukowców nr KAE/BMN/15/15.

1. WPROWDZENIE

W modelowaniu popytu powszechnie stosowane są modele ARIMA oraz regresja logistyczna. Briffaut i Lallement (2010) badają popyt w budownictwie za pomocą programowania dynamicznego. Klug (2011) stosuje symulację Monte Carlo w monitorowaniu popytu dla przedsiębiorstwa z branży motoryzacyjnej. Rippe i in. (1976) wykorzystują klasyczne modele szeregów czasowych w badaniu rynków przemysłowych.

Na popyt w przemyśle hutniczym wpływają m.in. takie czynniki, jak polityka gospodarcza i środowiskowa, zmiany koniunktury, działania konkurencji oraz działania marketingowe firmy. Niektóre z tych czynników obserwowane są jedynie pośrednio poprzez oddziaływanie na wielkość badanego zjawiska, a ocena ich wpływu jest subiektywna. Modele o charakterze deterministycznym nie dają możliwości uwzględniania zmiennych niejawnych. Przesłanki te potwierdzają zasadność badania popytu w przemyśle hutniczym za pomocą modelu ze zmiennymi ukrytymi.

Wprowadzone przez Bauma i Petriego (1966) ukryte modele łańcuchów Markowa stanowią klasę procesów stochastycznych służących do modelowania zjawisk o charakterze sekwencyjnym. Zjawisko przedstawione jest jako proces Markowa, dla którego obserwacja jest losową funkcją nieobserwowalnego stanu. Jednym z pierwszych zastosowań ukrytych łańcuchów Markowa było rozpoznawanie znaków. Z czasem zyskały na popularności w genetyce i biochemii, gdzie stosowane są m.in. w rozpoznawaniu mowy (Rabiner, Juang, 1991) czy modelowaniu sekwencji biologicznych. Za ich pomocą badano zjawiska przyrody (Zucchini, Guttorp, 1991), szacowano występowanie ataków epilepsji (Albert, 1991), analizowano wydarzenia polityczne (Schrodt, 2006). W literaturze dużo uwagi poświęca się analizie cykli koniunkturalnych oraz punktów zwrotnych z wykorzystaniem modeli o charakterze niedeterministycznym (m.in. Hamilton, 1990). Nguyen i Nguyen (2015) za pomocą modeli Markowa prognozują wielkość zmiennych makroekonomicznych wpływających na ceny akcji.

W artykule przedstawiono wyniki analizy popytu w przemyśle hutniczym, posługując się modelami Markowa mającymi niedeterministyczny charakter, które stanowią alternatywę dla powszechnie stosowanych zaawansowanych metod analizy szeregów czasowych. W literaturze brak przykładów stosowania tego typu modeli w prognozowaniu wskaźników przemysłu ciężkiego, będącego pod dużym wpływem niemierzalnych lub niejawnych czynników. Uwzględnienie tych oddziaływań jest ważne dla zwiększenia wiarygodności rezultatu.

Celem badania jest budowa ukrytego modelu łańcuchów Markowa, który odzwierciedli punkty zwrotne zapotrzebowania na wyroby hutnicze oraz umożliwi prognozę wielkości tego zapotrzebowania. Jest to bardzo istotne z punktu wi-

dzenia uczestników rynku planujących rozmiar produkcji, zapasów i zatrudnienie. W porównaniu z tradycyjnymi metodami prognozowania modele Markowa uwzględniają losowy charakter kształtowania się zjawiska, zamiast opierać się jedynie na historycznych danych.

Liczbę stanów w modelu Markowa wyodrębniono na podstawie teoretycznej analizy wartości szeregu. Najlepszy model ustalono na podstawie algorytmu Bauma-Welcha, zadając różne parametry początkowe. Rezultaty porównano z algorytmem *backward*. W celu weryfikacji rezultatu dokonano symulacji wartości dla 1000 szeregów na podstawie wyznaczonego modelu i sprawdzono ich własności. Wyniki porównano z wartościami empirycznymi. Następnie dokonano rekalkulacji parametrów modelu z wyłączeniem zbioru testowego oraz porównano prognozę krótko- i średniookresową z prognozą jednej z powszechnie stosowanych metod badania szeregów czasowych – ARIMA. ARIMA i modele Markowa opierają się na odmiennych założeniach, dlatego uzyskanie podobnego rezultatu jest jedną z metod walidacji modelu Markowa.

2. CHARAKTERYSTYKA UKRYTYCH MODELI ŁAŃCUCHÓW MARKOWA

Cappé i in. (2005) oraz Bernardelli (2013) formułują następującą definicję modeli Markowa na gruncie terminologii z dziedziny automatów skończonych[3]: S_X jest niepustym, skończonym k-elementowym zbiorem stanów ze stanem π, będącym stanem początkowym. Wówczas macierzą przejść jednorodnego łańcucha Markowa jest

$$P = \left[p_{ij}\right]_{i,j=1}^{k}, \tag{1}$$

gdzie $i, j = 1, \dots, k$ oraz $p_{ij} = P(X_t = j | X_{t-1})$. Określone w ten sposób p_{ij} jest prawdopodobieństwem przejścia ze stanu i do stanu j. Dodatkowo macierz przejść jest stochastyczna, co implikuje, że dla każdego i

$$\sum_{j=1}^{k} p_{ij} = 1. \tag{2}$$

Ukryte modele łańcuchów Markowa są procesami stochastycznymi, charakteryzowanymi przez jednorodny łańcuch Markowa o określonej liczbie stanów oraz losowe funkcje związane ze stanami. Stanowią one poszerzenie definicji łańcuchów Markowa o pewien alfabet S_Y, którego znaki są w danym stanie generowane z określonym prawdopodobieństwem. Łańcuch Markowa definiowany jest

[3] Automat skończony stanowi abstrakcyjny model zachowania systemu, dokonujący przejść między stanami w kolejnych iteracjach na podstawie tablicy przejść o wartościach dyskretnych (por. Mitchell, 1997).

jako jednorodny w czasie, gdy prawdopodobieństwo przejścia między stanami nie zależy od momentu t. Dany proces znajdujący się w stanie S_X w momencie t emituje symbol $i \in A$ z prawdopodobieństwem $e_i(x)$. W momencie $t+1$ przechodzi do stanu j z prawdopodobieństwem p_{ij}. Zakładamy też, że każdy stan emituje znak, zatem

$$\sum_{x \in A} e_i(x) = 1. \tag{3}$$

W każdym momencie t proces znajduje się w jednym stanie, a obserwacja generowana jest przez pewną losową funkcję. Kolejne stany łańcucha Markowa nie są obserwowalne, znane są jedynie realizacje procesu jako wyniki działania losowych funkcji. Definiujemy w ten sposób złożony proces $(X_t, Y_t)_{t=1}^{\infty}$, gdzie $(X_t)_{t=1}^{\infty}$ stanowi łańcuch Markowa, a $(Y_t)_{t=1}^{\infty}$ jest ciągiem obserwacji $y_t \in S_Y$. Prawdopodobieństwa emisji znaku z alfabetu $S_Y = 1,2,\dots,M$ zapisywane są w macierzy $\prod = [\pi_{xy}]$ o wymiarach $(N \times M)$, gdzie $x \in S_X | y \in S_Y$ (Pietrzykowski, Sałabun, 2014).

Cechą łańcucha Markowa jest zależność bieżącego stanu jedynie od stanu poprzedniego. Skrondal i Rabe-Hesketh (2004) mówią o konieczności rozróżnienia dwóch rodzajów zależności osiągnięcia danego stanu. Na podstawie badania zatrudnienia zamężnych kobiet w Stanach Zjednoczonych autorzy wskazują na większą skłonność podjęcia zatrudnienia przez kobiety, które pracowały w przeszłości, m.in. ze względu na nabytą wiedzę i doświadczenie. Jednak nie jest to jedyne wytłumaczenie i stan może być wywołany przez inne, nieoczywiste czynniki, niezwiązane z historią zatrudnienia, czyli przeszłym wystąpieniem tego stanu.

Dla k-elementowej przestrzeni stanów ukrytego łańcucha Markowa mówimy o k-stanowym modelu Markowa. W zależności od typu badanego szeregu liczbę stanów może stanowić np. średnia liczba obserwacji lub empiryczna obserwacja i interpretacja ekonomiczna stanów. Dla przykładu Bartolucci i inni (2009) w modelu oceny wydajności domów opieki społecznej za liczbę stanów przyjmują stan zdrowia pacjentów. Ścieżka kolejnych stanów wyznaczana jest na podstawie algorytmu Viterbiego, który oblicza najbardziej prawdopodobny ciąg na podstawie wartości obserwowanych w kolejnych okresach. Zwykle stosuje się modele dwu-, trzy- i czterostanowe ze względu na możliwość interpretacji rezultatu. Średni czas pozostawania w danym stanie określony jest wzorem:

$$v(i) = \frac{1}{1 - p_{i,i}}, \tag{4}$$

z wyjątkiem przypadku, gdy $p_{ij} = 1$. Stany i i j komunikują się, gdy możliwe jest przejście ze stanu i do j i odwrotnie. Jeśli wszystkie stany się komunikują, łańcuch nazywamy nieprzywiedlnym (Bernardelli, 2013).

Model ukrytych łańcuchów Markowa wymaga wyznaczenia wektora prawdopodobieństw stanu początkowego $\pi = \{\pi_i\}$. Przez $\pi_i = P(X_i = i)$ oznaczamy prawdopodobieństwo, że i jest stanem początkowym. $\pi = [\pi_i]$ to wektor o wymiarach $(1 \times N)$, który reprezentuje rozkład początkowy. Zwiernik (2005) podaje, że z tradycji modelowania ukrytych łańcuchów Markowa wynika założenie, że rozkład początkowy procesu jest rozkładem stacjonarnym. Rozkład prawdopodobieństwa $(\pi_i)_{i \in S_x}$ nazywamy stacjonarnym, gdy $\pi = \pi P$, tj. $\sum_{j \in S_x} \pi_j = 1$ i dla każdego j jest spełnione $\pi_j \geq 0$ oraz $\pi_j = \sum_i \pi_i p_{ij}$. Jak podkreśla Zwiernik (2005), założenie to nie odgrywa istotnej roli przy wystarczająco dużej liczbie obserwacji.

Zgodnie z powyższą notacją uporządkowana trójka (π, P, Π) stanowi łańcuch Markowa o N stanach i M znakach. Figielska (2011) zaznacza, że model określony jest poprzez rodzaj powiązań między stanami. W przypadku gdy przejście do danego stanu może nastąpić z każdego innego stanu, mamy do czynienia z modelem ergodycznym. W modelu Bakisa (tzw. model „od lewej do prawej") stany występują w określonej sekwencji, tzn. numer stanu wzrasta wraz ze zmianą t. Tego typu model właściwy jest dla modelowania sygnałów o własnościach zmieniających się w czasie. Nie jest możliwe przejście do stanu o numerze mniejszym od poprzedniego, a sekwencja zaczyna się w stanie 1.

W przypadku dyskretnych szeregów czasowych przyjmuje się, że kolejne obserwacje są generowane przez jednorodny proces Poissona (Zwiernik, 2005). Wówczas wyniki w następujących po sobie okresach są niezależnymi zmiennymi losowymi o wartości oczekiwanej równej wariancji, zgodnie z charakterystyką rozkładu Poissona. Spodziewamy się jednak, że występuje wzajemna zależność między obserwacjami, przeszłe obserwacje wpływają na bieżące, a wariancja może być wyższa od średniej. Zakłada się wówczas, że obserwacje są generowane przez odrębne procesy Poissona, o różnych średnich (λ_1 i λ_2 dla dwóch stanów), między którymi występuje proces przelicznikowy. λ_1 wybierane jest z prawdopodobieństwem δ_1, a λ_2 z prawdopodobieństwem $\delta_2 = 1 - \delta_1$. Uzyskana w ten sposób wariancja jest większa od średniej. Przyjmując, że procesem przełącznikowym jest łańcuch Markowa, a nie ciąg niezależnych zmiennych losowych, otrzymujemy wzajemną zależność między obserwacjami. Przedstawiony model wykorzystali Albert (1991) oraz Le i in. (1992) w analizie ataków padaczki.

Innym podejściem w analizowaniu rozkładów dyskretnych jest przełączanie rozkładów wielomianowych. Dla danej liczby obserwacji model wymaga oszacowania warunkowego rozkładu emisji dla N stanów, wyrażonego przez prawdopodobieństwo sukcesu p_i oraz $N^2 - N$ elementów macierzy przejścia leżących poza główną przekątną (Zwiernik, 2005). Proces Y_t o rozkładzie dwumianowym charakteryzują dwa parametry: liczba prób n oraz prawdopodobieństwo sukcesu w jednej próbie p, przy czym n nie musi być stałe w czasie. Wówczas p_i określa prawdopodobieństwo sukcesu związane ze zdarzeniem $\{X_i = i\}$. Crowder i inni (2005) stosują takie podejście w modelowaniu rynku obligacji.

Autorzy analizują sytuację na rynku inwestycyjnym dla różnych sektorów w stanie normalnego i podwyższonego ryzyka. Z kolei Azzalini i Bowman (1990) posługują się modelem opartym na rozkładzie wielomianowym w badaniu erupcji gejzera.

Ukryte modele Markowa rozkładów ciągłych znalazły zastosowanie w analizie cykli koniunkturalnych ze względu na brak możliwości bezpośredniego zmierzenia aktywności gospodarczej. Badania empiryczne wskazują na występowanie okresów ekspansji i recesji. Stan, w jakim znajduje się gospodarka, nie jest znany, można natomiast analizować pewne wielkości makroekonomiczne jako wskaźniki tej aktywności, np. PKB. Hamilton (1990) proponuje ukryty model Markowa badający zwroty cyklu dla Stanów Zjednoczonych na podstawie prawdopodobieństwa przełączania między dwoma łańcuchami.

3. PRZESŁANKI WYBORU MODELU ZE WZGLĘDU NA SPECYFIKĘ PRZEMYSŁU HUTNICZEGO

Trudno zidentyfikować wszystkie czynniki wpływu i skalę ich oddziaływania na popyt. W przypadku zjawisk nieobserwowalnych bezpośrednio istotną kwestią pozostaje sposób pomiaru zmiennej ukrytej. Niektóre zmienne można pośrednio wprowadzić do modelu za pomocą odpowiedniego indeksu, np. wskaźnika koniunktury, który pozwala na oszacowanie wielkości zjawiska. Modelowanie opiera się na zidentyfikowaniu związków między zmiennymi wskaźnikowymi. Zmienne objaśniające mogą być liniowymi kombinacjami ukrytych czynników. Wskaźnik klimatu koniunktury może np. odzwierciedlać przewidywania przedsiębiorców dotyczące kondycji rynku. Analiza zmiennych mierzalnych lub deklarowanych pozwala na wykorzystanie ich w modelu jako źródła informacji o czynnikach niemierzalnych pod warunkiem zidentyfikowania związku między nimi. Jednak nie wszystkie zmienne posiadają taki indeks, a wybór wskaźnika jest subiektywny. Wynika stąd potrzeba wykorzystania modeli klas ukrytych. Zastosowanie modeli niejawnych łańcuchów Markowa pozwala na oszacowanie skali wpływu zjawiska bez znajomości jego konkretnych wartości. Kolejne stany łańcucha Markowa nie są obserwowalne, znane są jedynie następujące realizacje procesu.

Wybór modelu podyktowany jest specyfiką przemysłu ciężkiego. Produkcja hutnicza przeznaczona jest dla przemysłu energetycznego, metalurgicznego, maszynowego i okrętowego. Wielkość popytu zależy od zapotrzebowania w gałęziach powiązanych, co powoduje, że kondycja sektora jest mocno sprzężona z ogólną koniunkturą gospodarczą (Hidalgo Gonzalez, Kamiński, 2011). Odbiorcami rzadko są indywidualni konsumenci. Wytwarza się przede wszystkim części składowe, a popyt powiązany jest z zainteresowaniem odbiorców produktami końcowymi. Mała elastyczność cenowa popytu wynika ze specjalizacji. Wśród pozacenowych czynników wpływających na zainteresowanie ofertą Urbaniak (1999) wymienia: jakość, przestrzeganie wymogów ochrony środowiska, koniunkturę gospodarczą, politykę gospodarczą, działania konkurencji, działania

marketingowe, wdrażanie przedsięwzięć infrastrukturalnych o dużej skali. Popyt można pobudzić poprzez uczestnictwo w targach lub reklamę w czasopismach branżowych. Znaczenie mają kompetencje przedstawicieli handlowych, udzielane upusty i możliwość negocjowania cen. Poza tym istotna jest lokalizacja przedsiębiorstwa, pozwalająca na zmniejszenie kosztów dostaw. Ważną cechą rynku jest konieczność dostarczenia produktu o określonych parametrach oraz złożony cykl produkcji, co wiąże się z wysokimi kosztami. Obecność na rynkach międzynarodowych pozwala na dywersyfikację rynków zbytu. Zdarza się jednak, że przedsiębiorstwa spotykają się z preferencyjnym traktowaniem lokalnych producentów lub z niesprzyjającą polityką gospodarczą goszczącego kraju.

Struktura rynku opiera się na kilku dużych przedsiębiorstwach oraz kilkudziesięciu średniej i małej wielkości zakładach i zbliżona jest do oligopolu, zgodnie z definicją przedstawioną przez Tirole (1988). Na rynku działa stosunkowo niewielu producentów, a decyzje w sprawie cen i wielkości produkcji kształtują się w zależności od poczynań pozostałych. Przedsiębiorstwa ustalają optymalną dla siebie strategię działania. Wzrost cen w sektorze może nastąpić np. wskutek wzrostu kosztów energii.

Liczba potencjalnych nabywców jest względnie stała. Skupieni są oni w określonych regionach. Odbiorcami wyrobów są także dystrybutorzy i jednostki wewnętrzne przedsiębiorstwa, np. działy marketingu oraz badań i rozwoju, pracujące nad promocją lub ulepszaniem technologii. Przemysł charakteryzuje się niską elastycznością zwiększania mocy wytwórczych. Występują korzyści skali. Z tego powodu wymagany jest długi horyzont prognozy. W długim okresie firmy mogą dostosować się do nowych warunków poprzez zwiększenie zatrudnienia lub rozbudowę zakładów. Rozmiar produkcji podyktowany jest maksymalizacją zysku, a nie wielkością popytu. Koszt krańcowy początkowo jest wysoki, co wiąże się z alokacją środków w maszyny i technologie, ale stopniowo maleje.

Dla pełnego obrazu uwarunkowań przemysłu hutniczego należy również uwzględnić kontekst historyczny. Tradycje hutnictwa sięgają XVI w. Przez lata udoskonalano proces technologiczny, a wyroby znajdowały coraz więcej zastosowań. Rozwój w okresie powojennym miał związek z odbudową zniszczeń. W czasie transformacji gospodarki w latach 90. XX w. państwowe przedsiębiorstwa napotykały trudności z dostosowaniem się do nowych wymogów (Nowara, Szarzec, 2004). W wielu przypadkach proces restrukturyzacji i możliwość samodzielnego funkcjonowania bez pomocy państwa były możliwe dzięki przejęciu przez zagranicznego inwestora, a przedsiębiorstwa stawały się częścią ponadnarodowych holdingów. Kwilinski (2018) podkreśla, że w gospodarce opartej na wiedzy zakłady produkcyjne stoją przed wyzwaniami związanymi z koniecznością wspierania przedsiębiorczości, nawiązania współpracy z centrami badawczymi, inwestowania w kapitał ludzki, wdrażania infrastruktury informatycznej. Wprowadzanie tych zmian jest procesem długotrwałym, wymagającym rewizji sposobu zarządzania i struktur organizacyjnych, doskonalenia systemów infor-

matycznych i logistyki, ale prowadzi do poprawy konkurencyjności firmy (Rachwał, 2008).

W ostatnich latach obserwujemy zmienne zapotrzebowanie na wyroby polskich hut, a jego skala zależy od rodzaju produkcji i kondycji rynku odbiorcy. Przemysł motoryzacyjny, postrzegany jako jeden z głównych odbiorców, notuje okresy wzrostu i spadku, a przedsiębiorstwa hutnicze muszą dostosowywać się do zmieniających się trendów i zainteresowania lżejszymi samochodami o mniejszych silnikach, których części wytwarzane są w kuźniach[4]. Dobre perspektywy ma przed sobą budownictwo związane z ochroną środowiska, transportem i energetyką, za sprawą środków unijnych przeznaczonych na inwestycje tego typu. Prognozuje się również dynamiczny rozwój przemysłu energetycznego w zakresie alternatywnych form pozyskania energii, m.in. energii ze źródeł odnawialnych, energii jądrowej, energetyki wiatrowej (Hutnicza Izba Przemysłowo-Handlowa, 2017). Przemysł stoczniowy boryka się z licznymi trudnościami, chociaż nadal pozostaje kluczowy ze względu na handel drogą morską (Instytut Studiów Wschodnich, 2018).

Modele klas ukrytych pozwalają oszacować sumaryczny wpływ wymienionych czynników na popyt bez konieczności ich identyfikacji.

4. ESTYMACJA PARAMETRÓW MODELU

Estymacji parametrów dokonuje się na podstawie algorytmu Bauma-Welcha (Baum, Eagon, 1967), algorytmu prefiksowo-sufiksowego (Rabinier, 1989; Jurafsky, Martin, 2008) oraz metod gradientowych (Jelinek, 1976). Algorytm Bauma-Welcha polega na znalezieniu w kolejnych iteracjach lokalnego maksimum funkcji wiarygodności $\mathcal{L}(\theta)$. Opiera się na trzech formułach reestymacyjnych dla parametrów modelu i wymaga wyprowadzenia estymatorów NWW dla parametrów rozkładu początkowego, prawdopodobieństw przejść i parametrów warunkowych gęstości emisji (Zwiernik, 2005). Algorytm prefiksowo-sufiksowy wymaga policzenia prawdopodobieństwa prefiksowego i sufiksowego dla wszystkich stanów N i okresów t. Jest to metoda ogólna polegająca na obliczeniu funkcji wiarygodności $\mathcal{L}(\theta) = P\left(y^{(T)}\middle|\theta\right)$ wystąpienia danej sekwencji obserwacji $y^{(T)}$ przy zadanych parametrach modelu (Zwiernik, 2005). Prawdopodobieństwa prefiksowe $\alpha_t(i)$ i sufiksowe $\beta_t(i)$ wyznacza się jako:

$$\alpha_t(i) = \mathrm{P}(Y_1 = y_1, \dots, Y_T = y_t, X_T = i|\theta), \tag{5}$$

$$\beta_t(i) = \mathrm{P}(Y_{t+1} = y_{t+1}, \dots, Y_T = y_t, X_T = i|\theta). \tag{6}$$

[4] http://www.euroforge.org/industry-portrait/trends-in-europe.html (dostęp: 1.10.2018).

Prawdopodobieństwa dla kolejnych stanów i okresów wyznaczane są rekurencyjnie, a wartości początkowe wyznaczone są dla przyjętych parametrów modelu jako:

$$\alpha_1(i) = \mathrm{P}(X_1 = i)P(Y_1 = y_1, X_1 = i|\theta) = \delta_i \pi_{iy_1}, \tag{7}$$

$$\beta_T(i) = 1. \tag{8}$$

Metody oparte na gradientach bazują na wielkości funkcji w punktach oraz informacji o pochodnych. Lokalne minimum znajduje się poprzez wyznaczanie kolejnych kierunków poszukiwań w oparciu o gradient funkcji celu, w punkcie osiągniętym w poprzedniej iteracji.

Nie ma gwarancji znalezienia optymalnych parametrów, a jakość oszacowania zależy od przyjętych wartości początkowych. Otrzymane rozwiązanie może być jedynie optimum lokalnym dla zadanych wartości startowych (Figielska, 2012; Bernardelli, 2013). W tej sytuacji należy rozważyć powtórzenie algorytmu dla różnych wartości początkowych i wybór optymalnego rozwiązania na podstawie wartości kryteriów informacyjnych.

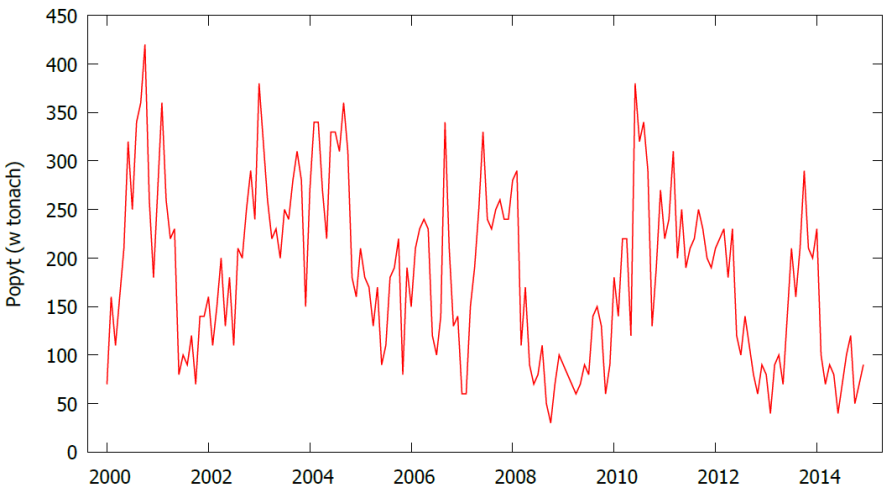
5. MODEL

Badanie przeprowadzono dla lat 2000–2014 z wykorzystaniem danych jednego z producentów wyrobów hutniczych w Polsce. Firma jest jednym z większych przedstawicieli rynku, którego strukturę można określić jako kilkanaście dużych i średnich zakładów oraz kilkadziesiąt małych kuźni. Produkcja przeznaczona jest głównie na rynek krajowy i rynki europejskie. Przedsiębiorstwo przeszło typową dla tego typu zakładów drogę związaną z transformacją polskiej gospodarki po upadku komunizmu. W latach 90. XX w. zmagało się z dostosowaniem do gospodarki rynkowej oraz silną konkurencją na skutek otwarcia granic. Proces dostosowawczy był długotrwały, ale zakończył się pomyślnie przy udziale zagranicznego kapitału. Obecnie przedsiębiorstwo jest częścią dużej europejskiej grupy. Inwestuje w nowe technologie i rozwój kadr, wprowadza innowacje w zakresie zarządzania, dba o ciągłe usprawnianie procesu produkcji z poszanowaniem środowiska, przeznacza środki na promocję i umacnianie wizerunku firmy, a także stosuje strategię upustów dla stałych odbiorców. Firma stara się również dywersyfikować rodzaj produkcji i rynki zbytu, co skutkuje mniejszą wrażliwością na załamania na rynkach odbiorców. W obliczu silnej konkurencji przedsiębiorstw z Azji kładzie duży nacisk na jakość wyrobów i uzyskiwanie certyfikatów. W ostatnich latach ograniczyła produkcję dla przemysłu okrętowego, który boryka się z trudnościami.

Szereg na rys. 1, przedstawiający zapotrzebowanie na wyroby dla przemysłu maszynowego, charakteryzuje się znacznymi wahaniami. Fluktuacje mogą być wynikiem zmian polityki środowiskowej, koniunktury gospodarczej, kondycji powiązanych przemysłów lub być efektem sezonowym.

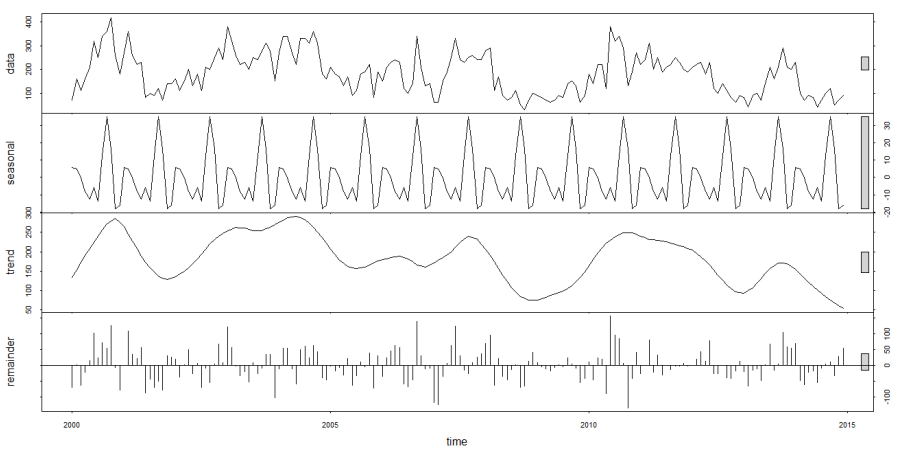
Na podstawie testu ADF odrzucono hipotezę o występowaniu pierwiastka jednostkowego przy $p = 0{,}0282$. Hipotezę o normalności rozkładu odrzucono na podstawie testu Shapiro-Wilka $(p = 0{,}0003)$ i testu Jarque'a-Bery $(p = 0{,}03)$. Za pomocą procedury stl programu R wyznaczono trend, element sezonowy oraz element losowy (rys. 2). Rysunek 3 jest ilustracją skali wahań sezonowych i losowych względem właściwego szeregu. Przeprowadzone w programie Demetra testy sezonowości nie potwierdzają jej występowania. Biorąc to pod uwagę, a także rezultat procedury programu R, która wskazuje na niski udział wahań sezonowych względem oryginalnego szeregu, potraktowano składową sezonową oraz składową losową jako realizacje składowej obserwowalnej.

Rysunek 1. Miesięczny popyt na wyroby przeznaczone dla przemysłu maszynowego



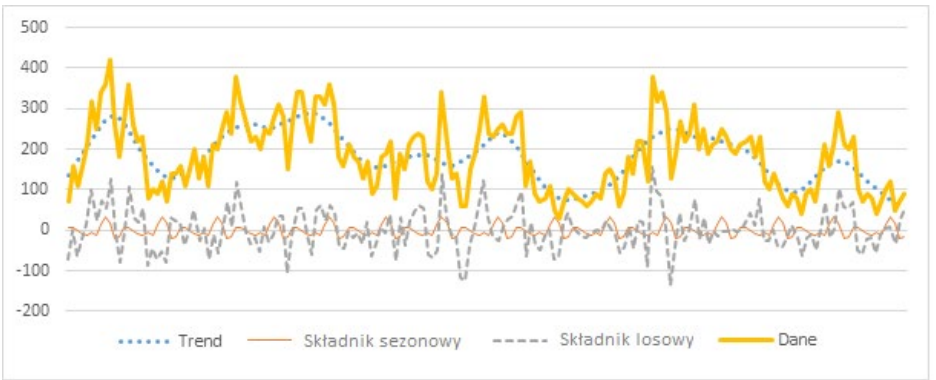
Ź r ó d ł o: obliczenia własne w programie Gretl.

Rysunek 2. Procedura stl



Ź r ó d ł o: obliczenia własne w programie R.

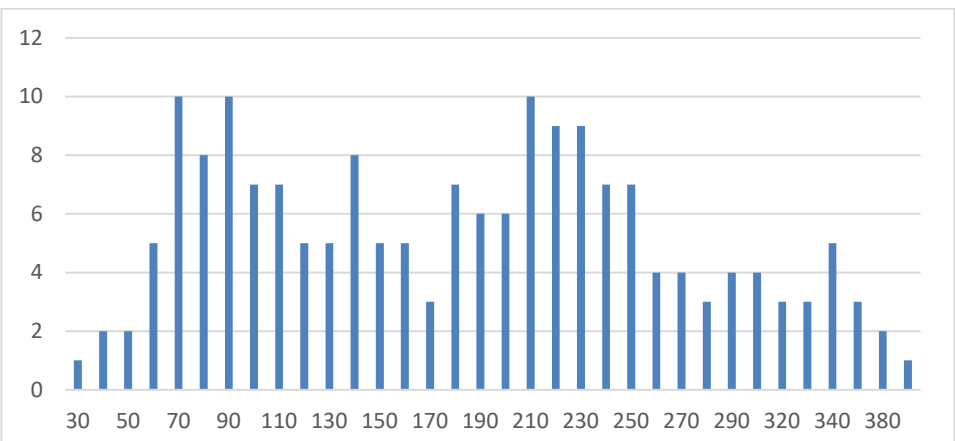
Rysunek 3. Skala wahań sezonowych i losowych względem oryginalnego szeregu



Ź r ó d ł o: obliczenia własne w programie JDemetra+.

Ze względu na mnogość czynników wpływu oraz brak miary, która umożliwiłaby ich bezpośrednie wprowadzenie do modelu, przyjęto, że stanowią one zmienne ukryte. Obserwujemy jedynie ich sumaryczny wpływ na badaną zmienną. Na podstawie wykresu szeregu nie można wnioskować o liczbie stanów ukrytego modelu łańcuchów Markowa. Przed przystąpieniem do obliczenia częstości występowania poszczególnych wartości i wyznaczenia liczby stanów zmodyfikowano szereg, zaokrąglając wszystkie wartości do 5 ton. Uzyskano w ten sposób 34 unikalne obserwacje spośród 180 obserwacji ogółem. Rysunek 4 ilustruje liczność poszczególnych obserwacji.

Rysunek 4. Histogram częstości obserwacji



Ź r ó d ł o: obliczenia własne w programie Excel.

W badanym szeregu możemy wyodrębnić trzy podgrupy, dla których częstość obserwacji przebiega zgodnie z rozkładem normalnym. Jako granice dla podgrup wyznaczono wartości 160 i 280. Tablica 1 przedstawia średnie i wariancje w podgrupach. Odchylenia standardowe w podgrupach są znacząco mniejsze niż dla wszystkich obserwacji. W ostatniej podgrupie, dla której wartości szeregu są najwyższe, występuje jedynie 25 obserwacji, co nie pozwala na wiarygodne zbadanie własności szeregu. W pozostałych dwóch grupach nie ma podstaw do odrzucenia hipotezy o normalności rozkładów (na podstawie testu Jarque'a-Bery).

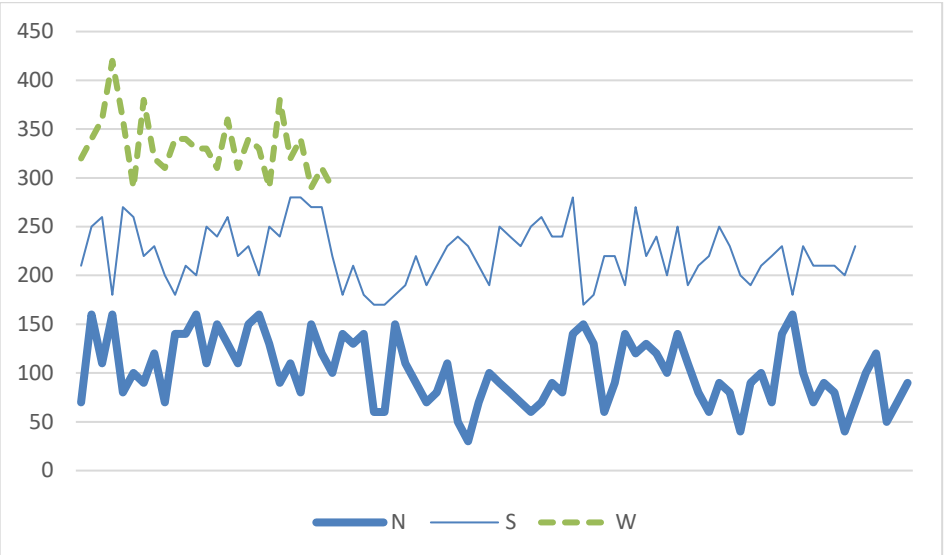
TABLICA 1. ŚREDNIE I WARIANCJE W GRUPACH

Wyszczególnienie	$T(n=180)$	$N(n=80)$	$S(n=75)$	$W(n=25)$
Średnia	183,9	101,8	222,0	332,4
Odchylenie standardowe	87,7	33,5	29,5	31,9

Ź r ó d ł o: opracowanie własne.

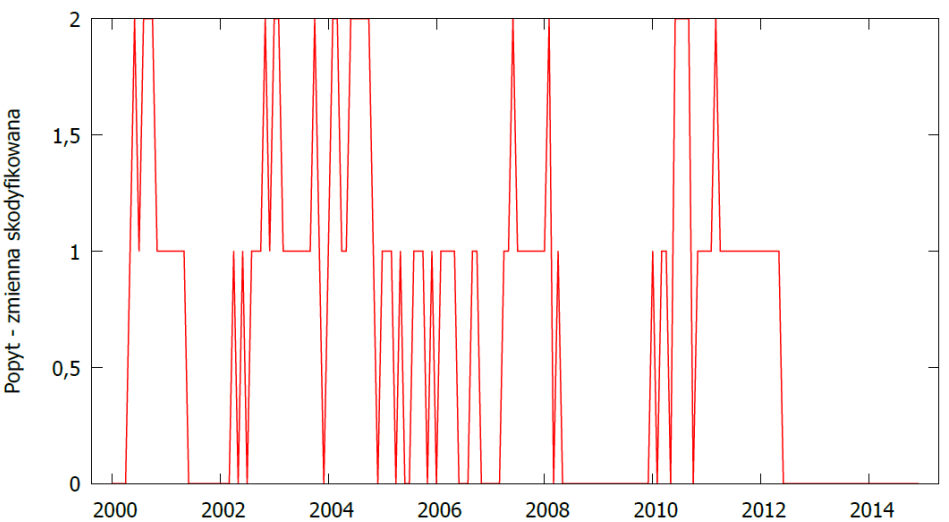
Rysunek 5 jest ilustracją szeregów dla wartości niskich, umiarkowanych i wysokich. Uwzględniając to, że w grupie trzeciej występuje mniejsza liczba obserwacji, można zauważyć, że kształtują się one podobnie. Analiza histogramu oraz różnice wielkości odchyleń standardowych w poszczególnych podgrupach mogą wskazywać na występowanie trzech stanów. Model wymaga zbudowania macierzy emisji o wymiarach $N \times M$, należałoby zatem wyznaczyć $N^2 + N(M-2)$ warunkowych prawdopodobieństw. W przypadku $N = 3$ i $M = 34$ staje się to nieefektywne. W celu ograniczenia liczby parametrów przyjęto model dla szeregu o rozkładzie wielomianowym. Dla danej liczby obserwacji warunkowy rozkład emisji dla trzech stanów jest wyznaczony przez prawdopodobieństwo sukcesu p_i. Należy zatem wyznaczyć prawdopodobieństwa sukcesu p_i dla trzech stanów oraz sześciu elementów macierzy przejścia leżących poza przekątną. X_t jest łańcuchem Markowa o wartościach z trójelementowej przestrzeni stanów $S_X = \{N, S, W\}$. Stany określono umownie jako odpowiadające niskim, średnim i wysokim wartościom popytu. Przez $\pi_i = P(X_i = i)$ oznaczamy prawdopodobieństwo, że stanem początkowym jest stan i. Wektor $\pi = [\pi_1 \pi_2, \pi_3]$ jest rozkładem początkowym. Początkowe prawdopodobieństwa znalezienia się procesu w stanach N, S, W wynoszą odpowiednio 0,444, 0,417 i 0,319. Są one równoważne udziałowi niskich, umiarkowanych i wysokich wartości popytu w badanym szeregu.

Rysunek 5. Prezentacja graficzna szeregów dla wartości niskich, umiarkowanych i wysokich



Ź r ó d ł o: obliczenia własne w programie Excel.

Rysunek 6. Wielkość popytu – zmienna skodyfikowana



Ź r ó d ł o: obliczenia własne w programie Gretl.

Celem konstrukcji modelu jest wyznaczenie trendu kształtowania popytu oraz punktów zwrotnych, nie zaś konkretnych wartości zjawiska. Chcemy określić, kiedy wystąpiły okresy wzmożonego i niskiego popytu. Dla uproszczenia modelu skodyfikowano dane, tworząc zbiór obserwacji $S_Y = \{0,1,2\}$, alfabet składa się zatem z trzech znaków. Klasyfikację oparto na tej samej podstawie co w przypadku wyznaczania stanów. Wartość 0 określa niskie, a 2 wysokie wartości popytu. Rysunek 6 jest ilustracją zmian popytu dla zmiennej skodyfikowanej. Proces można opisać następująco: po kilku okresach występowania niskich wartości popytu następują okresy przejściowe wartości umiarkowanych, naprzemiennie z krótkimi okresami, kiedy notujemy wartości wysokie. Wyznaczono początkowe składniki macierzy przejścia między stanami. Korzystając z procedur języka R i zadając różne parametry początkowe, ustalono optymalne parametry macierzy przejścia i macierzy emisji (tabl. 2 i 3).

TABLICA 2. MACIERZ PRZEJŚCIA

Stan	N	S	W
N	0,786	0,138	0,076
S	0,211	0,647	0,141
W	0,043	0,046	0,911

Źródło: opracowanie własne.

TABLICA 3. MACIERZ EMISJI

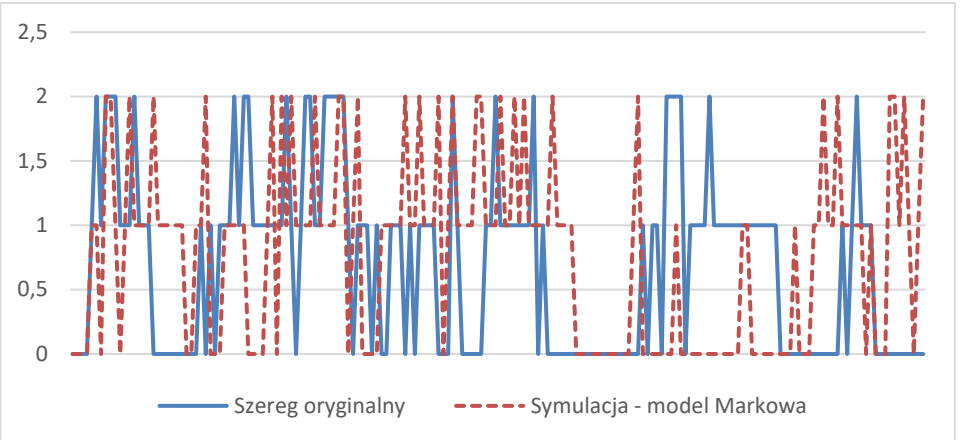
Stan	0	1	2
N	0,91	0,08	0,00
S	0,79	0,20	0,01
W	0,06	0,68	0,26

Źródło: opracowanie własne.

Zauważmy, że wszystkie stany się komunikują. Macierz przejścia wskazuje, że rzadko zachodzi przejście ze stanu W do stanu N i odwrotnie. Średni czas przebywania w stanie N jest równy 1/(1-0,786), czyli 4,7 okresu, w stanie S nieco ponad 3 okresy, a w stanie W – 11 okresów.

Symulację wartości szeregu przeprowadzono zgodnie z wyznaczonymi parametrami modelu Markowa. Rysunek 7 przedstawia szereg oryginalny i symulowany. Obserwacje z symulacji modelu Markowa częściowo pokrywają się z oryginalnym szeregiem.

Rysunek 7. Wartości rzeczywiste i prognozowane



Ź r ó d ł o: obliczenia własne w programie Excel.

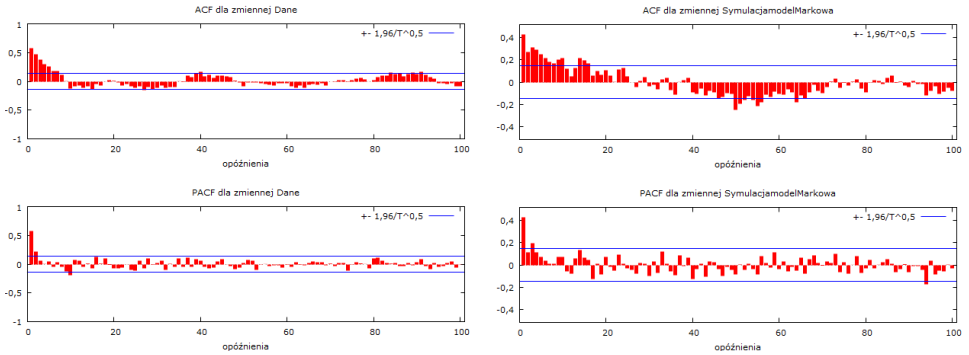
Tablica 4 zawiera wartości funkcji autokorelacji (ACF) i autokorelacji cząstkowej (PACF) właściwego szeregu. Funkcja ACF stopniowo wygasa, a liczba niezerowych wartości jest równa 7. Funkcja autokorelacji cząstkowej przyjmuje istotne wartości do drugiego opóźnienia. Dla wygenerowanego na podstawie parametrów modelu Markowa szeregu funkcja ACF kształtuje się nieco odmiennie, tzn. pozostaje istotna dla większej liczby opóźnień (por. rysunek 8).

TABLICA 4. FUNKCJE ACF I PACF ORYGINALNEGO SZEREGU

Opóźnienie	ACF	PACF	Ljung-Box Q	*p*
1	0,5750*	0,5750*	60,5108	[0,000]
2	0,4788*	0,2214*	102,7079	[0,000]
3	0,3827*	0,0644	129,8083	[0,000]
4	0,2978*	0,0043	146,3172	[0,000]
5	0,2662*	0,0496	159,5830	[0,000]
6	0,1858	–0,0415	166,0808	[0,000]
7	0,1768	0,0390	172,0027	[0,000]
8	0,1078	–0,0520	174,2141	[0,000]
9	0,0046	–0,1310	174,2183	[0,000]
10	–0,1211	–0,1884	177,0461	[0,000]
11	–0,0847	0,0735	178,4373	[0,000]
12	–0,0744	0,0541	179,5177	[0,000]
13	–0,1095	–0,0459	181,8702	[0,000]
14	–0,0914	0,0109	183,5173	[0,000]
15	–0,1412	–0,0670	187,4777	[0,000]

* wartości większe niż $\pm 1{,}96/T^{0,5}$
Ź r ó d ł o: opracowanie własne.

Rysunek 8. Funkcje ACF i PACF szeregu oryginalnego i symulowanego



Ź r ó d ł o: obliczenia własne w programie Gretl.

Funkcje ACF i PACF mogą wskazywać, że odpowiedni do modelowania szeregu będzie łańcuch Markowa drugiego rzędu. Prawdopodobieństwa początkowe znalezienia się procesu w danym stanie ustalono tak jak poprzednio. W tabl. 5 i 6 przedstawiono oszacowane parametry macierzy przejścia i macierzy emisji.

TABLICA 5. MACIERZ PRZEJŚCIA MODELU DLA ŁAŃCUCHA MARKOWA DRUGIEGO RZĘDU

S t a n	*N*	*S*	*W*
N	0,97	0,02	0,00
S	0,50	0,33	0,03
W	0,08	0,63	0,18

Ź r ó d ł o: opracowanie własne.

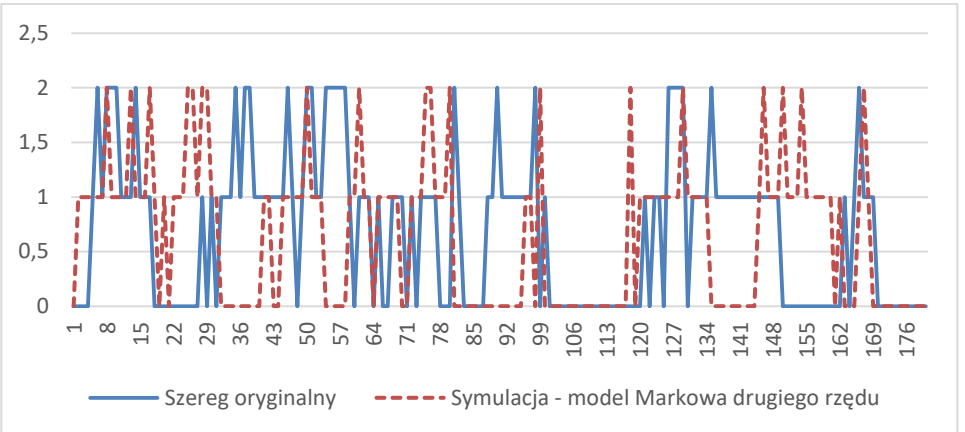
TABLICA 6. MACIERZ EMISJI MODELU DLA ŁAŃCUCHA MARKOWA DRUGIEGO RZĘDU

S t a n	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)	(2,0)	(2,1)	(2,2)
N	0,67	0,08	0,01	0,09	0,09	0,04	0,00	0,01	0,01
S	0,09	0,12	0,01	0,09	0,43	0,09	0,01	0,09	0,05
W	0,00	0,08	0,00	0,12	0,28	0,08	0,04	0,16	0,24

Ź r ó d ł o: opracowanie własne.

W wyniku symulacji, polegającej na wygenerowaniu 180 obserwacji spełniających założenia procesu Markowa, o wyznaczonych poprzednio parametrach, uzyskano szereg przedstawiony na rys. 9.

Rysunek 9. Porównanie oryginalnego szeregu z symulacją – model drugiego rzędu

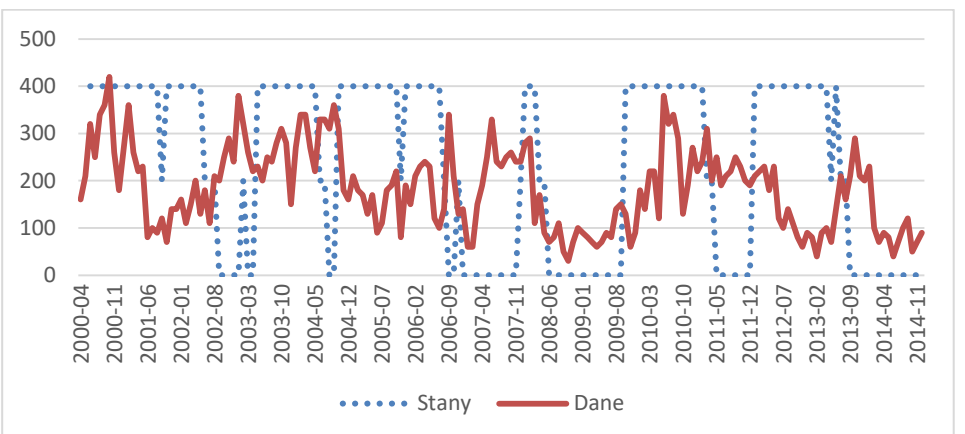


Ź r ó d ł o: obliczenia własne w programie Excel.

Tym razem funkcje ACF i PACF kształtują się jak dla oryginalnego szeregu (rys. 11). Model drugiego rzędu uwzględnia występującą korelację.

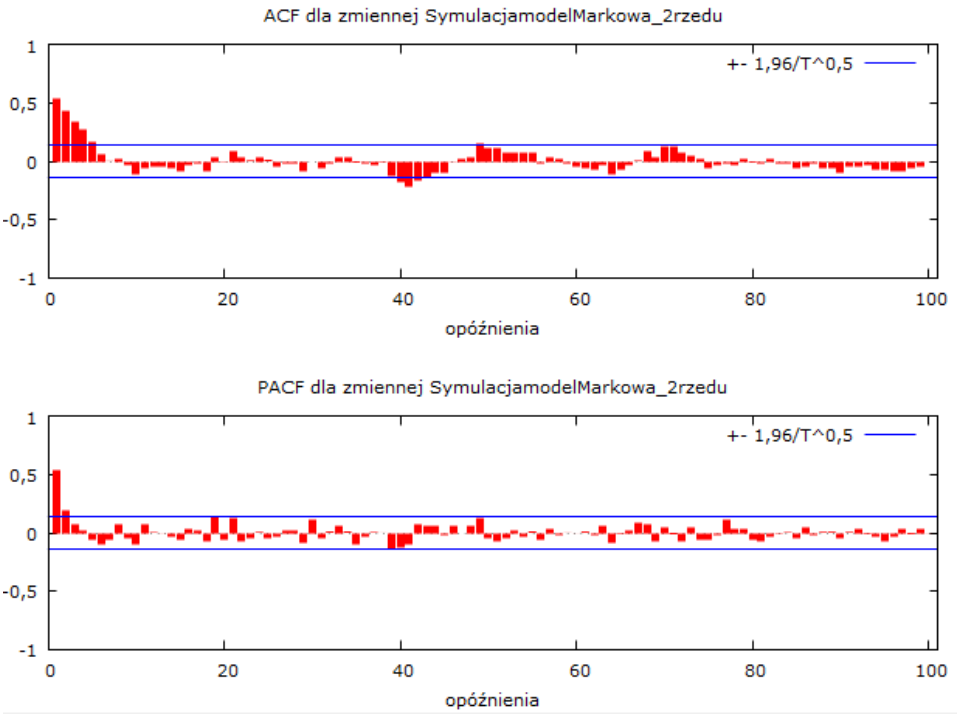
Do określenia prawdopodobnego przebiegu stanów wykorzystano algorytm Viterbiego. Rysunek 10 przedstawia, jak kształtują się zmiany stanów względem kolejnych obserwacji szeregu. Obserwujemy stan spadku popytu dla okresu od kwietnia 2008 r. do października 2009 r. oraz w 2014 r. Brak jest jednak odniesienia do innych niż badany szereg danych referencyjnych, które wskazałyby, czy wyznaczone stany pokrywają się z obserwacjami empirycznymi.

Rysunek 10. Algorytm Viterbiego – przebieg stanów względem rzeczywistych obserwacji



Ź r ó d ł o: obliczenia własne w programie Excel.

Rysunek 11. Funkcje ACF i PACF dla modelu drugiego rzędu



Ź r ó d ł o: obliczenia własne w programie Gretl.

Rezultat uzyskany przez wygenerowanie na podstawie modelu pojedynczego szeregu może mieć charakter przypadkowy. W celu oceny jego jakości przeprowadzono symulację polegającą na wygenerowaniu 1000 szeregów dla modelu Markowa drugiego rzędu. Tablica 7 przedstawia średnie różnice między wartościami rzeczywistymi a wartościami prognozowanymi.

TABLICA 7. ŚREDNIE RÓŻNICE MIĘDZY WARTOŚCIAMI RZECZYWISTYMI A PROGNOZOWANYMI

Różnica	Liczba obserwacji	Skumulowana liczba obserwacji	%	Skumulowany %
0	81	81	45	45
–1	40	121	22	67
+1	39	160	22	89
–2	12	172	7	96
+2	8	180	4	100

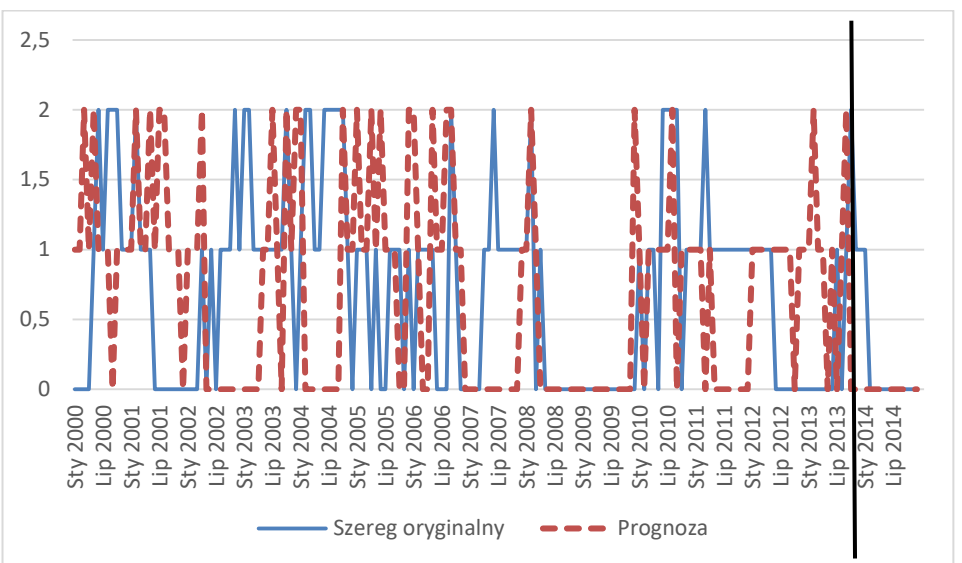
Ź r ó d ł o: opracowanie własne.

Wartość 0 oznacza zbieżność obu wartości. Wartości –1 i –2 oznaczają niedoszacowanie modelu względem wartości rzeczywistych i pomyłkę o odpowiednio jedną lub dwie wartości. Wartości +1 i +2 oznaczają przeszacowanie modelu. Najmniej pożądane jest niedoszacowanie lub przeszacowanie o dwie wartości, które odnotowano w 11% przypadków. Trafną prognozę uzyskano średnio dla 45% przypadków. Różnice rozkładają się symetrycznie, brak tendencji do przeszacowania lub niedoszacowania. Niedoszacowanie lub przeszacowanie modelu o jedną wartość również jest niepożądane, bo wartości szeregu są skodyfikowane, a rozpiętości przedziałów wynoszą odpowiednio 130 t, 120 t i 140 t dla kolejnych grup. Szeregi generowane przez model Markowa w dużym stopniu korelują z szeregiem referencyjnym.

6. PROGNOZA

Dokonano rekalkulacji modelu dla skróconego szeregu, pozostawiając 12 ostatnich obserwacji jako zbiór testowy. Przyjęto te same wartości prawdopodobieństw początkowych. Wartości parametrów macierzy przejścia i macierzy emisji uległy nieznacznej modyfikacji w porównaniu z tabl. 5 i 6. Rysunek 12 ilustruje prognozę dla 12-miesięcznego okresu. Wartości zbioru testowego oddzielono pionową linią. Tylko dla pierwszej obserwacji ze zbioru testowego wartość prognozy odbiega od wartości rzeczywistej – w tym przypadku popyt zaklasyfikowano jako umiarkowany, a według prognozy jego wartość jest niska. Dla pozostałych obserwacji wartości prognozowane pokrywają się z rzeczywistymi (tabl. 8).

Rysunek 12. Prognoza dla 12 miesięcy



Ź r ó d ł o: obliczenia własne w programie Gretl.

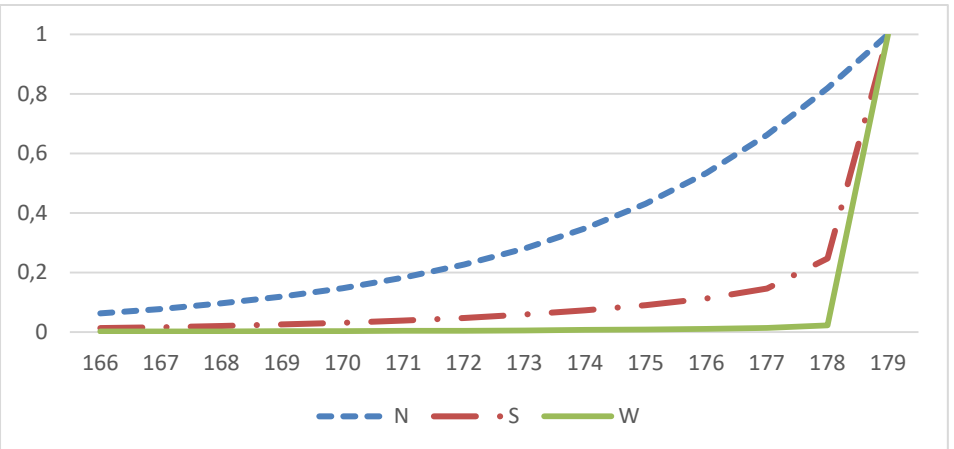
TABLICA 8. WARTOŚCI PROGNOZOWANE I RZECZYWISTE

Nr obserwacji	169	170	171	172	173	174	175	176	177	178	179	180
Szereg oryginalny	1	0	0	0	0	0	0	0	0	0	0	0
Wartości prognozowane	0	0	0	0	0	0	0	0	0	0	0	0

Ź r ó d ł o: opracowanie własne.

Na rys. 13 zaprezentowano obliczenia algorytmu *backward* dla 12 ostatnich obserwacji. Wskazuje on na prawdopodobieństwo wystąpienia danej sekwencji obserwacji $[k+1, \ldots, n]$ pod warunkiem wystąpienia stanu X w momencie k. Zgodnie z oczekiwaniami dany ciąg obserwacji jest najbardziej prawdopodobny w stanie niskiego popytu.

Rysunek 13. Wartości wyznaczone w kolejnych krokach algorytmu *backward*

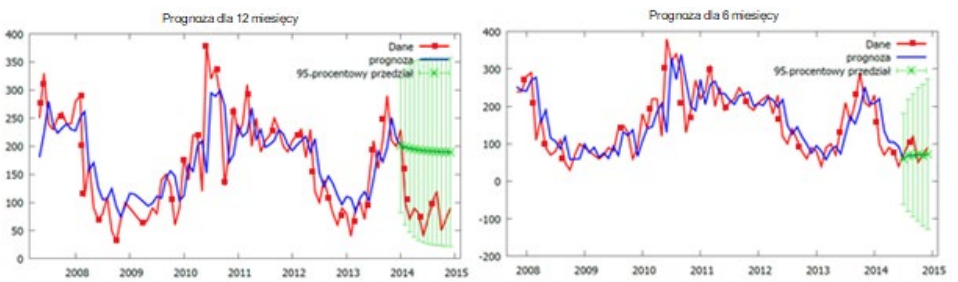


Ź r ó d ł o: obliczenia własne w programie Excel.

W prognozowaniu szeregów czasowych często stosuje się modele typu ARIMA, dlatego prognozę dla ARIMA porównano z prognozą modelu Markowa. Z szeregu wyłączono 12 i 6 ostatnich obserwacji, aby dokonać prognozy długo- i średniookresowej (rys. 14). W obu przypadkach najlepszy rezultat osiągnięto dla modelu ARIMA (1,0,1) (na podstawie najniższych wartości kryteriów informacyjnych i szerokości przedziału ufności). Analiza korelogramu nie wskazała na występowanie sezonowości. W przypadku prognozy długookresowej wyłączono wszystkie obserwacje z 2014 r., kiedy nastąpił znaczny spadek wartości popytu. Prognoza modelu ARIMA nie odzwierciedla tego spadku. W prognozowanym okresie obserwujemy niski popyt i według przyjętej klasyfikacji wartości oznaczono by jako 0. Wartości prognozy oscylują natomiast wokół 200, zatem zostałyby zaklasyfikowane jako 1 – umiarkowany popyt. Dużo lepszy wynik uzyskano w przypadku prognozy średniookresowej, gdzie wartości empiryczne

kształtują się podobnie jak wartości rzeczywiste, uwzględniając wygładzanie szeregu przez model ARIMA. Zarówno wartości prognozowane, jak i rzeczywiste zostałyby zaklasyfikowane jako 0 – niski popyt. W przypadku prognozy długookresowej lepszy rezultat uzyskano by zatem dla modelu Markowa, który uwzględnił duży spadek popytu w analizowanym okresie.

Rysunek 14. Model ARIMA (1,0,1) – prognoza długo- i średniookresowa



Ź r ó d ł o: obliczenia własne w programie Gretl.

7. PODSUMOWANIE

Zaproponowany model ukrytych łańcuchów Markowa drugiego rzędu może być alternatywą dla tradycyjnych metod analizy szeregów czasowych (m.in. metod wyrównywania wykładniczego czy ARIMA). Pozwala na uwzględnienie zmiennych niemierzalnych bądź ukrytych. Dobrze sprawdził się w prognozowaniu dłuższych okresów obniżonego popytu. Rezultat ten należałoby potwierdzić dla dłuższego szeregu lub innego zestawu danych. W przypadku prognozy średniookresowej sprawdził się lepiej niż model ARIMA, w którym zabrakło informacji o długotrwałym obniżeniu poziomu popytu związanym z potencjalną zmianą stanu w prognozowanym okresie.

Ograniczeniem modelu jest kodyfikacja wartości szeregu, która powoduje utratę części informacji. Wyznaczona prognoza dostarcza informacji na temat kształtowania się trendu zamiast wyznaczenia konkretnych wartości popytu. Możemy ją jednak potraktować jako wskazówkę co do punktów zwrotnych koniunktury.

Wprowadzenie modelu trójstanowego pozwoliło na uwzględnienie tzw. stanu przejściowego jako sygnału o nadchodzącej zmianie pomiędzy fazami wzrostu i spadku. Podobna próba z modelem dwustanowym nie przyniosłaby wystarczająco użytecznej informacji dla badanego przedsiębiorstwa.

LITERATURA

Albert P. S., (1991), A Two-State Markov Mixture Model for a Time Series of Epileptic Seizure Counts, *Biometrics*, 47(4), 1371–1381. DOI: 10.2307/2532392.

Azzalini A., Bowman A. W., (1990), A Look at Some Data on the Old Faithful Geyser, *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 39(3), 357–365. DOI: 10.2307/2347385.

Bartolucci F., Lupparelli M., Montanari G. E., (2009), Latent Markov Model for Longitudinal Binary Data: An Application to the Performance Evaluation of Nursing Homes, *The Annals of Applied Statistics*, 3(2), 611–636. DOI: 10.1214/08-AOAS230.

Baum L. E., Eagon J. A., (1967), An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model of Ecology, *Bulletin of the American Mathematicians Society*, 73(3), 360–363. DOI: 10.1090/S0002-9904-1967-11751-8.

Baum L. E., Petrie T., (1966), Statistical Interference for Probabilistic Functions of Finite State Markov Chains, *The Annals of Mathematical Statistics*, 37(6), 1554–1563. DOI: 10.1214/aoms/ /1177699147.

Bernardelli M., (2013), Nieklasyczne modele Markowa w analizie cykli koniunktury gospodarczej w Polsce, *Roczniki Kolegium Analiz Ekonomicznych*, 30, 59–74.

Briffaut J. P., Lallement P., (2010), Volatility Forecasting of Market Demand as Aids for Planning Manufacturing Activities, *Service Science & Management*, 3, 383–389. DOI: 10.1109/ICCIE.2009. 5223926.

Cappé O., Moulines E., Rydén T., (2005), *Inference in Hidden Markov Models*, Springer.

Crowder M., Davis M., Giampieri G., (2005), Analysis of Default Data Using Hidden Markov Model of Default Interaction, *Quantitative Finance*, 5(1), 27–34. DOI: 10.1080/14697680500039951.

Figielska E., (2011), Ewolucyjne metody uczenia ukrytych modeli Markowa, *Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyk*i, 5, 63–74.

Hamilton J. D., (1990), Analysis of Time Series Subject to Changes in Regime, *Journal of Econometrics*, 45(1–2), 39–70. DOI: 10.1016/0304-4076(90)90093-9.

Hidalgo Gonzalez I., Kamiński J., (2011), The Iron and Steel Industry: A Global Market Perspective, *Gospodarka Surowcami Mineralnymi*, 27(3), 5–28.

Hutnicza Izba Przemysłowo-Handlowa, (2017), *Polski przemysł stalowy 2017. Pobrane z:* http://www.hiph.org//ANALIZY_RAPORTY/pliki/PPS-2017.pdf.

http://www.hiph.org//ANALIZY_RAPORTY/pliki/PPS-2017.pdf.

Instytut Studiów Wschodnich, (2008), *Polski przemysł stoczniowy – stan obecny, perspektywy, zagrożenia*, Forum Ekonomiczne, Warszawa.

Jelinek F. (1976), Continuous Speech Recognition by Statistical Methods, *Proceedings of the IEEE*, 64(4), 532–556. DOI: 10.1109/PROC.1976.10159.

Jurafsky D., Martin J. H., (2008), *Speech and Language Recognition*, Prentice Hall.

Klug F., (2011), Automotive Supply Chain Logistics: Container Demand Planning using Monte Carlo Simulation, *International Journal of Automotive Technology and Management*, 11(3), 254–268. DOI: 10.1504/IJATM.2011.040871.

Kwilinski A., (2018), Mechanism of formation of industrial enterprise development strategy in the information economy*, Virtual Economics*, 1(1), 7–25. DOI: 10.34021/ve.2018.01.01(1).

Le N.D., Leroux B. G., Puterman M. L., (1992), Reading Reaction: Exact Likelihood Evaluation in a Markov Mixture Model for Time Series of Seizure Counts, *Biometrics*, 48(1), 317–323. DOI: 10.2307/2532758.

Mitchell T. M., (1997), *Machine Learning*, McGraw-Hill Corporation, New York.

Nguyen N., Nguyen D., (2015), Hidden Markov Model for Stock Selection, *Risks*, 3(4), 455–473. DOI: 10.3390/risks3040455.

Nowara W., Szarzec K., (2004), Skutki procesów upadłościowych i układowych przedsiębiorstw w Polsce w latach 1990–2002, w: A. Manikowski, A. Psyk (red.), *Unifikacja gospodarek europejskich: szanse i zagrożenia*, Wydawnictwo Naukowe Wydziału Zarządzania Uniwersytetu Warszawskiego, Warszawa.

Pietrzykowski M., Sałabun W., (2014), Applications of Hidden Markov Model: State-of-the-Art, *International Journal of Computer Technology and Applications*, 5(4), 1384–1391.

Rabiner L. R., (1989), Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, 77(2), 257–286. DOI: 10.1109/5.18626.

Rabiner L. R., Juang B. H., (1991), Hidden Markov Models for Speech Recognition, *Technometrics*, 33(3), 251–272. DOI: 10.2307/1268779.

Rachwał T., (2008), Problematyka badawcza funkcjonowania przedsiębiorstw przemysłowych, w: Z. Zioło, M. Rachwał (red.), *Problematyka badawcza geografii przemysłu*, 11, 53–85. Wydawnictwo Naukowe AP, Kraków.

Rippe R., Wilkinson W., Morrison D., (1976), Industrial Market Forecasting with Anticipations Data, *Management Science*, 22(6), 639–651. DOI: 10.1287/mnsc.22.6.639.

Schrodt P. A., (2006), Forecasting Conflict in the Balkans using Hidden Markov Models, w: R. Trapple (ed.), *Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention*, Springer, 161–184.

Skrondal A., Rabe-Hesketh S., (2004), *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman and Hall/CRC, Boca Raton.

Tirole J., (1988), *The Theory of Industrial Organization*, MIT Press, Cambridge.

Urbaniak M., (1999), *Marketing przemysłowy*, Wydawnictwo Infor, Warszawa.

Zucchini W., Guttorp P., (1991), A Hidden Markov Model for Spacetime Precipitation, *Water Resources Research*, 27(28), 1917–1923. DOI: 10.1029/91WR01403.

Zwiernik P. W., (2005), *Wstęp do ukrytych modeli Markowa i metody Bauma-Welcha.* Pobrane z: http://www.mimuw.edu.pl/~pzwiernik/docs/hmm.pdf.

DOI: 10.5604/01.3001.0014.0952

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 4 – 2019

Kamila TRZCIŃSKA[1]

Aproksymacja rozkładu dochodów ludności Polski za pomocą modeli Daguma i Zengi

Streszczenie. *Celem pracy jest analiza dochodów ludności Polski za pomocą rozkładów Daguma i Zengi. Rozkład Zengi zaproponowany w 2010 r. jest nowym, konkurencyjnym rozkładem, który nie był dotychczas stosowany do badania płac i dochodów w Polsce. W artykule przedstawiono rozkłady Daguma i Zengi oraz wyniki aproksymacji rozkładów płac pochodzących z badania budżetów gospodarstw domowych ludności Polski w 2014 r. Obliczone miary zgodności dopasowania rozkładów teoretycznych do empirycznych wskazują, że rozkład Zengi lepiej opisuje rozkład płac ludności Polski niż – uważany dotychczas za jeden z najlepszych – rozkład Daguma.*

Słowa kluczowe: rozkład dochodów, model Daguma, model Zengi, metody estymacji

Approximation of distribution of Poland's population's income according to Dagum and Zenga models

Abstract. *The aim of this paper is to analyse the income of the population of Poland using the Dagum and the Zenga distributions. The Zenga distribution, introduced in 2010, is a new distribution which has not been yet applied to analysing wages and income in Poland. The paper presents both the Dagum and Zenga distributions, as well as the results of the approximations of wage distributions drawn from the 2014 household budget survey. The calculated degree of compliance of the theoretical distribution with the empirical one demonstrates that the Zenga distribution describes the distribution of the income of the Polish population more reliably than the Dagum distribution, which so far has been regarded as one of the best.*

Keywords: income distribution, Dagum distribution, Zenga distribution, estimation methods

JEL Classification: C1, C10, C13, C15

[1] Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny, Katedra Metod Statystycznych, ul. Rewolucji 1905 r. 41/43, 90-214 Łódź, Polska, e-mail: kamila.trzcinska@uni.lodz.pl, ORCID: https://orcid.org/0000-0002-4714-4074.

1. WPROWDZENIE

Pierwszą usystematyzowaną teorię dotyczącą rozkładu i nierównomierności dochodów przedstawił Pareto w przełomowej pracy z 1897 r. Następnie pojawiły się prace dotyczące teoretycznych rozkładów płac i dochodów. March (1898) zaproponował wykorzystanie rozkładu gamma, a Edgeworth (1898) – zastosowanie rozkładu normalnego jako funkcji generującej rozkłady dochodów. Podejście to przyczyniło się do stworzenia m.in. przez D'Addaria (1939) pierwszego systemu funkcji gęstości opisujących rozkłady płac i dochodów. Zagadnienie rozwoju badań dotyczących dochodów i bogactwa na przestrzeni lat poruszają Domański i Jędrzejczak (2005). Obecnie badania rozkładów dochodów skupiają się głównie na próbach dopasowania rozkładów teoretycznych do empirycznego rozkładu płac i dochodów w różnych przekrojach oraz na analizie nierównomierności tych rozkładów.

Istnieje wiele teoretycznych rozkładów dochodów, które zostały sformułowane na podstawie obserwacji empirycznych (rozkład Pareta, rozkład Daguma) lub też jako wynik pewnego procesu stochastycznego (rozkład Champernowne'a). Przegląd większości z nich przedstawiono w pracy Kleibera i Kotza (2003). Dagum (1990) określił kryteria, które powinny spełniać teoretyczne modele rozkładów dochodów. Do najważniejszych z nich należą:

- zbieżność modelu do rozkładu Pareta typu I dla wysokich dochodów;
- istnienie małej liczby skończonych momentów;
- ekonomiczna interpretacja parametrów;
- możliwość dopasowania modelu do dwóch rodzajów rozkładów dochodów (w zależności od wartości parametrów powinien generować rozkłady jednomodalne lub zeromodalne).

Wiele rozkładów stosowanych w analizie płac i dochodów wybrano ad hoc, ze względu na ich kształt (rozkład gamma, rozkład beta, krzywą Pearsona V i VI typu). Empiryczne rozkłady bardzo często cechują się bowiem jednomodalnością, prawostronną asymetrią oraz dodatnią kurtozą. Są to prawidłowości, które zaobserwowano niezależnie od kraju oraz czasu. Do opisywania dochodów empirycznych poszukuje się takich rozkładów teoretycznych, które przy odpowiedniej parametryzacji odzwierciedlają te własności.

Zastosowanie rozkładów teoretycznych do analizy płac i dochodów jest uzasadnione z wielu powodów. Podejście takie umożliwia wyrównywanie nieregularności w obserwowanych rozkładach, które powstały w wyniku gromadzenia danych. Ponadto ułatwia analizę rozkładów, ponieważ wszystkie charakterystyki rozkładów teoretycznych, w tym także współczynniki nierównomierności, mogą być wyrażone za pomocą tych samych parametrów. Znajomość parametrów rozkładu może być też wykorzystana do porównania tych samych rozkładów w różnych okresach oraz do konstrukcji modeli rozkładów dochodów. Dane, którymi dysponujemy, pochodzą na ogół z badań częściowych, dlatego znajomość funkcji gęstości lub dystrybuanty teoretycznej ułatwia proces estymacji

miar nierównomierności dla zbiorowości generalnej. Wielu badaczy koncentruje się na problemach porównań rozkładów dochodów w czasie i przestrzeni. Stosowanie rozkładów teoretycznych do opisu empirycznych rozkładów dochodów stwarza możliwość prognozowania.

Do aproksymacji rozkładów dochodów ludności Polski stosowano różne rozkłady teoretyczne, np. rozkłady Pareta, logarytmiczno-normalny, gamma, Daguma i wiele innych (Wiśniewski, 1934; Vielrose, 1960; Lange, 1967; Wąsik, 1967; Kordos, 1968, 1973; Domański, 1973; Kot, 1999, 2000). Zastosowaniem dwuparametrowego rozkładu log-normalnego zajmowali się Kordos (1990) oraz Jagielski i Kutner (2010).

Badania nad rozkładami dochodów i płac w Polsce wskazują, że do aproksymacji empirycznych rozkładów płac i dochodów mogą być z powodzeniem stosowane rozkłady teoretyczne trójparametrowe – w szczególności rozkłady Singha-Maddali i Daguma dobrze dopasowują się do danych empirycznych w różnych przekrojach (Łukasiewicz, Orłowski, 2004; Ostasiewicz, 2013; Salamaga, 2016). Rozkład Daguma, należący do krzywych Burra III typu (Kleiber, Kotz, 2003), jest najczęściej stosowany nie tylko w Polsce. Jest on nieco lepiej dopasowany od rozkładu Singha-Maddali, należącego do krzywych Burra XII typu (Jędrzejczak, 1999), a oprócz tego jest zbieżny z krzywą Pareta dla wysokich grup dochodowych.

W 2010 r. Zenga zaproponował trójparametrowy rozkład oparty na uciętym rozkładzie Pareta przedstawionym w pracy Polisicchia (2008), który bardzo dobrze aproksymuje rozkłady dochodów wielu krajów. Rozkład ten nie był dotąd stosowany do analizy rozkładów ludności Polski. Funkcja gęstości Zengi, dzięki swoim własnościom, łatwo dopasowuje się do różnego rodzaju dochodów.

Do opisu rozkładów dochodów ludności Polski stosowano również czteroparametrowy rozkład GB2. Badania prowadzone przez Banduriana i in. (2003), Dastrupa i in. (2007) oraz Brzezińskiego (2013) potwierdzają bardzo dobre dopasowanie tego rozkładu do danych empirycznych.

W niniejszym artykule przedstawiono porównanie rozkładu Daguma, uważanego do tej pory za rozkład trójparametrowy najlepiej opisujący rozkłady empiryczne w Polsce, z rozkładem Zengi, który pozwala na bezpośrednią interpretację parametrów.

2. TEORETYCZNE ROZKŁADY DOCHODÓW

2.1. ROZKŁAD DAGUMA

Rozkład Daguma (Dagum, 1977) powstał na podstawie obserwacji empirycznych i jest jednym z najczęściej stosowanych do opisu rozkładów płac i dochodów. Rozkład ten może być zarówno jednomodalny, jak i zeromodalny. W przypadku wysokich wartości dochodów rozkład Daguma jest zbieżny z rozkładem Pareta, dzięki czemu posiada niewielką liczbę skończonych momentów, co sprawia, że charakteryzuje się wysoką zgodnością z danymi empirycznymi. Do

opisu rozkładu dochodów w Polsce został po raz pierwszy zastosowany przez Jędrzejczak (1990). W artykule zostanie rozważony rozkład Daguma typu I (Kleiber, Kotz, 2003). Funkcja gęstości rozkładu Daguma ma postać:

$$f(x: a; b; p) = \frac{apx^{ap-1}}{b^{ap}\left[1 + \left(\frac{x}{b}\right)^{a}\right]^{p+1}}, x > 0, \tag{1}$$

gdzie $a, b, p > 0$.

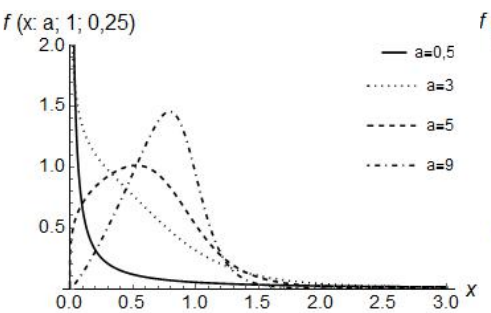
Natomiast dystrybuanta rozkładu Daguma jest następująca:

$$F(x: a; b; p) = \left[1 + \left(\frac{x}{b}\right)^{-a}\right]^{-p}, x > 0, \tag{2}$$

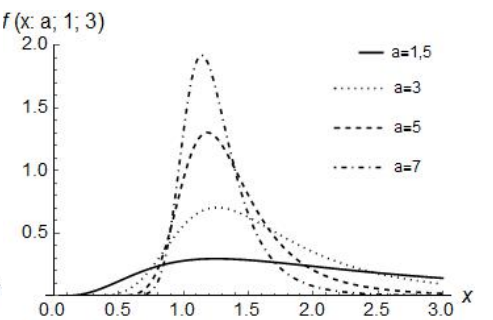
gdzie $a, b, p > 0$. Parametr b jest parametrem skali, natomiast a oraz p są parametrami kształtu.

Rysunek 1. Funkcja gęstości rozkładu Daguma

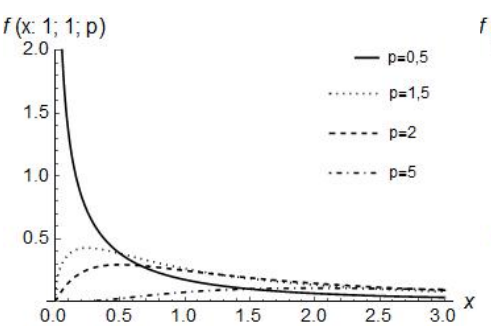
$f(a: 1; 0{,}25)$ dla $b = 1$ oraz $p = 0{,}25$



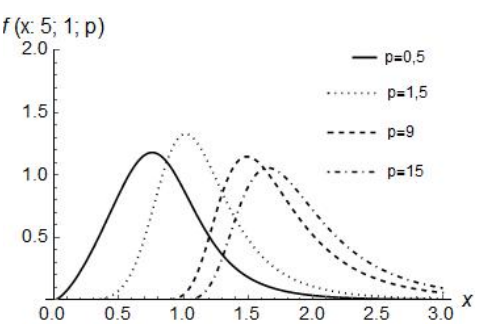
$f(a: 1; 3)$ dla $b = 1$ oraz $p = 3$



$f(x: 1; 1; p)$ dla $a = 1$ oraz $b = 1$



$f(x: 5; 1; p)$ dla $a = 5$ oraz $b = 1$



Ź r ó d ł o: opracowanie własne.

2.2. ROZKŁAD ZENGI

W 2010 r. Zenga skonstruował trzyparametrową funkcję gęstości $f(x{:}\,\mu;\alpha;\theta)$, $(\mu > 0;\ \alpha > 0;\theta > 0)$ dla nieujemnych wartości dochodów (Zenga, 2010; Zenga i in., 2010a, 2010b), która powstała w oparciu o rozkład Pareta przedstawiony przez Polisicchia (2008):

$$v(x{:}\,\mu;k) = \begin{cases} \dfrac{\sqrt{\mu}}{2}k^{0,5}(1-k)^{-1}x^{-1,5} & \text{dla} \quad \mu k \le x \le \dfrac{\mu}{k};\quad \mu > 0,\quad 0 < k < 1 \\ 0 & \text{dla pozostałych przypadków,} \end{cases} \tag{3}$$

oraz rozkład beta

$$g(k{:}\,\alpha;\theta) = \begin{cases} \dfrac{k^{\alpha-1}(1-k)^{\theta-1}}{\beta(\alpha;\theta)} & \text{dla} \quad 0 < k < 1;\quad \theta > 0, \alpha > 0 \\ 0 & \text{dla pozostałych przypadków,} \end{cases} \tag{4}$$

gdzie $B(\alpha;\theta)$ jest funkcją beta.

Funkcja gęstości rozkładu Zengi ma postać:

$$\begin{aligned} & f(x{:}\,\mu;\alpha;\theta) \\ & = \int_0^1 v(x{:}\,\mu,k)g(k{:}\,\alpha,\theta)dk \\ & = \begin{cases} \dfrac{1}{2\mu B(\alpha;\theta)}\left(\dfrac{x}{\mu}\right)^{-1,5}\displaystyle\int_0^{\frac{x}{\mu}} k^{\alpha-0,5}(1-k)^{\theta-2}dk & \text{dla} \quad 0 < x < \mu \\ \dfrac{1}{2\mu B(\alpha;\theta)}\left(\dfrac{\mu}{x}\right)^{1,5}\displaystyle\int_0^{\frac{\mu}{x}} k^{\alpha-0,5}(1-k)^{\theta-2}dk & \text{dla} \quad \mu < x. \end{cases} \end{aligned} \tag{5}$$

W przypadku gdy parametr $\theta > 0$, wówczas dystrybuanta rozkładu Zengi jest następująca:

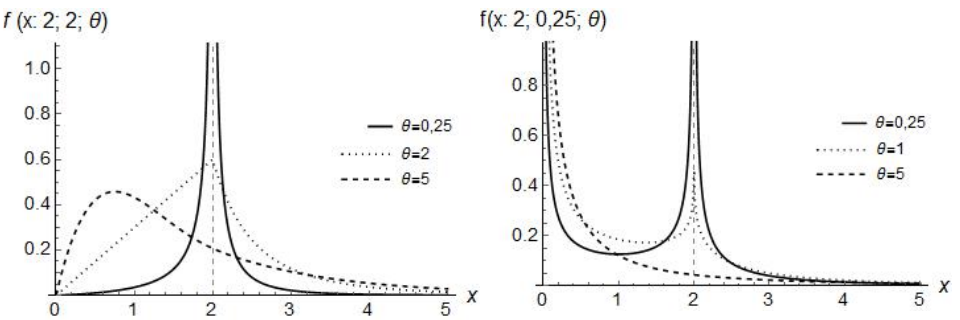
$$F(x{:}\,\mu;\alpha;\theta) = \begin{cases} \dfrac{1}{B(\alpha,\theta)}\displaystyle\sum_{i=1}^{\infty}\left\{IB\left(\frac{x}{\mu};\alpha+i-1;\theta\right) - \left(\frac{\mu}{x}\right)^{0,5} IB\left(\frac{x}{\mu};\alpha+i-0{,}5;\theta\right)\right\} & \text{dla} \quad 0 < x \le \mu \\ 1 - \dfrac{1}{B(\alpha,\theta)}\displaystyle\sum_{i=1}^{\infty}\left\{\left(\frac{\mu}{x}\right)^{0,5} IB\left(\frac{\mu}{x};\alpha+i-0{,}5;\theta\right) - IB\left(\frac{\mu}{x};\alpha+i;\theta\right)\right\} & \text{dla} \quad \mu < x, \end{cases} \tag{6}$$

gdzie $IB(x;\alpha;\theta)$ jest niekompletną funkcją beta. Ponadto $F(\mu{:}\,\mu;\alpha;\theta) \ge \frac{1}{2}$, dla $\alpha > 0;\ \theta > 0$, dlatego też rozkład ma dodatnią asymetrię.

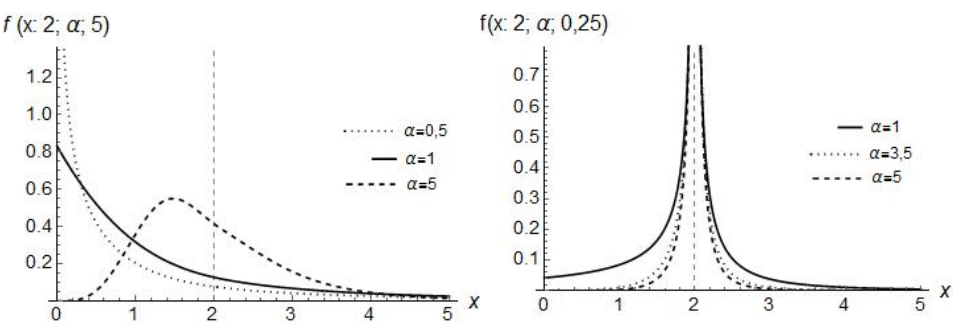
Łatwo zauważyć, że dla wszystkich wartości parametrów $\mu > 0$; $\alpha > 0$; $\theta > 0$ wartość oczekiwana rozkładu Zengi wynosi μ. Liczba skończonych momentów rozkładu jest określona jako mniejsza niż $\alpha + 1$, tak więc dla typowych wartości tego parametru 2 lub 3 rozkład Zengi posiada 3 lub 4 momenty. Funkcja gęstości rozkładu Zengi przyjmuje różne kształty, bardziej różnorodne niż tradycyjne trójparametrowe modele rozkładów dochodów, jak np. rozkład Daguma (por. rys. 1 i rys. 2), a cecha ta pozwala na dobre dopasowanie także w przypadku niskich grup dochodowych.

Rysunek 2. Funkcja gęstości rozkładu Zengi

$f(x; 2; 2; \theta)$ dla $\mu = 2$ oraz $\alpha = 2$ $f(x; 2; 0{,}25; \theta)$ dla $\mu = 2$ oraz $\alpha = 0{,}25$



$f(x; 2; \alpha; 5)$ dla $\mu = 2$ oraz $\theta = 5$ $f(x; 2; \alpha; 0{,}25)$ dla $\mu = 2$ oraz $\theta = 0{,}25$



Ź r ó d ł o: opracowanie własne.

Parametr μ jest parametrem skali rozkładu Zengi, natomiast α i θ są parametrami kształtu. W pracy zastosujemy trójparametrową funkcję gęstości Zengi do opisu empirycznego rozkładu dochodów ludności Polski na podstawie danych indywidualnych dochodu rozporządzalnego netto w gospodarstwach domowych w 2014 r. Źródłem danych są badania budżetów gospodarstw domowych prowadzone przez Główny Urząd Statystyczny.

3. METODY ESTYMACJI PARAMETRÓW ROZKŁADU ZENGI

Do estymacji parametrów funkcji Zengi $f(x:\mu;\alpha;\theta)$ użyjemy m.in. metod zaproponowanych przez D'Addaria (1934, 1939) oraz metody momentów. Przyjmujemy, że parametr $\mu = E(X)$ jest estymowany przez średnią z próby $\bar{x}$. Wówczas znalezienie parametrów funkcji Zengi α oraz θ polega na rozwiązaniu odpowiednich układów równań (Zenga i in., 2010a). Przyjmijmy następujące oznaczenia dla każdego z pięciu wariantów metody D'Addaria przedstawionych przez Zengę (2010):

- metoda 1: $F(\hat{m}:\bar{x},\alpha,\theta) = 0{,}5$ oraz $F(\bar{x}:\bar{x},\alpha,\theta) = \hat{F}(\bar{x})$, gdzie $\hat{m}$ jest empiryczną medianą, a $\hat{F}(\bar{x})$ – empiryczną dystrybuantą obliczoną dla średniej;
- metoda 2: $F(\hat{m}:\bar{x},\alpha,\theta) = 0{,}5$ oraz $P(\bar{x};\alpha;\theta) = \hat{P}$, gdzie $P(\bar{x};\alpha;\theta) = 2F(1:1;\alpha;\theta) - 1$ jest indeksem Pietry, a $\hat{P} = \frac{1}{2\bar{x}n}\sum_{i=1}^{n}|x_i - \bar{x}|$ – empirycznym indeksem Pietry;
- metoda 3: $F(\hat{m}:\bar{x},\alpha,\theta) = 0{,}5$ oraz $Var(\bar{x},\alpha,\theta) = m_2$, gdzie m_2 jest wariancją oraz $Var(\bar{x},\alpha,\theta) = \frac{\mu^2\theta(\theta+1)}{3(\alpha-1)(\alpha+\theta)}$;
- metoda 4: $F(\bar{x}:\bar{x},\alpha,\theta) = \hat{F}(\bar{x})$ oraz $Var(\bar{x},\alpha,\theta) = m_2$;
- metoda 5: $F(\hat{m}:\bar{x},\alpha,\theta) = 0{,}5$ oraz $A(\bar{x};\alpha;\theta) = \hat{A}(\bar{x})$, gdzie $A(\bar{x};\alpha;\theta) = 1 - \left\{\frac{1-F(1:1;\alpha;\theta)}{F(1:1;\alpha;\theta)}\right\}^2$ jest indeksem punktowym Zengi (2007) dla $\bar{x}$ oraz $\hat{A}(\bar{x}) = \frac{\sum_{x_i\le\bar{x}}x_i}{\hat{F}(\bar{x})} \times \frac{1-\hat{F}(\bar{x})}{\sum_{x_i>\bar{x}}x_i}$.

Zauważmy, że w celu wyznaczenia parametrów funkcji Zengi wszystkie z pięciu metod nakładają po trzy ograniczenia. Narzucając jedynie dwa ograniczenia – $\mu = \bar{x}$ oraz: $F(\hat{m}:\bar{x},\alpha,\theta) = 0{,}5$, $F(\bar{x}:\bar{x},\alpha,\theta) = \hat{F}(\bar{x})$, $P(\bar{x};\alpha;\theta) = \hat{P}$, $A(\bar{x};\alpha;\theta) = \hat{A}(\bar{x})$, otrzymujemy inwariantną metodę D'Addaria.

Zenga i in. (2010c) otrzymali analityczną metodę estymacji parametrów funkcji gęstości modelu Zengi poprzez rozwiązanie następującego układu równań:

$$\begin{cases} \mu = \bar{x} \\ \dfrac{\bar{x}^2\theta(\theta+1)}{3(\alpha-1)(\alpha+\theta)} = m_2 \\ \dfrac{3\bar{x}m_2(\theta+3)(\theta+2)}{5(\alpha+\theta+1)(\alpha-2)} = m_3. \end{cases}$$

Lewe strony równań oznaczają momenty teoretyczne, a $\bar{x}, m_2$ oraz m_3 są odpowiednio średnią arytmetyczną, wariancją oraz trzecim momentem centralnym

z próby. Aby układ równań miał rozwiązanie, trzeci moment centralny musi być skończony, czyli $\alpha > 2$. Wówczas estymowane parametry mają postać:

$$\hat{\mu} = \bar{x},$$

$$\hat{\theta} = \frac{-\left[\frac{1}{3}\frac{\bar{x}^2}{m_2} - 3\frac{\bar{x}m_2}{m_3} - 1\right] + \sqrt{\left[\frac{1}{3}\frac{\bar{x}^2}{m_2} - 3\frac{\bar{x}m_2}{m_3} - 1\right]^2 + 4\left[\frac{1}{3}\frac{\bar{x}^2}{m_2} - \frac{3}{5}\bar{x}\frac{m_2}{m_3}\right]\left[\frac{18}{5}\frac{\bar{x}m_2}{m_3} + 2\right]}}{2\left[\frac{1}{3}\frac{\bar{x}^2}{m_2} - \frac{3}{5}\bar{x}\frac{m_2}{m_3}\right]},$$

$$\hat{\alpha} = \frac{-\left(\hat{\theta} - 1\right) + \sqrt{\left(\hat{\theta} - 1\right)^2 + 4\left[\frac{1}{3}\frac{\bar{x}^2}{m_2}\hat{\theta}\left(\hat{\theta} + 1\right) + \hat{\theta}\right]}}{2}.$$

Do estymacji parametrów rozkładu Daguma stosowane są różne metody, które zostały omówione w pracach Domańskiego i Jędrzejczak (1998) oraz Jędrzejczak i in. (2018, 2019). W celu określenia stopnia zgodności rozkładów empirycznych z teoretycznymi wykorzystuje się rozbieżności pomiędzy empirycznymi wartościami różnych charakterystyk statystycznych rozkładu średniej arytmetycznej oraz odpowiednich indeksów (np. Giniego, Pietry, Theila).

W niniejszym artykule zgodność dopasowania rozważanych rozkładów została oceniona również za pomocą wskaźnika podobieństwa struktur, indeksu Mortara A_1, indeksu Pearsona kwadratów różnic częstości względnych A_2 oraz zmodyfikowanego indeksu kwadratów różnic częstości względnych A_2'. Indeksy te mają postać:

$$A_1 = \frac{1}{n}\sum_{j=1}^{s}\left|n_j - \hat{n}_j\right|, \tag{7}$$

$$A_2 = \sqrt{\frac{1}{n}\sum_{j=1}^{s}\frac{(n_j - \hat{n}_j)^2}{\hat{n}_j}}, \tag{8}$$

$$A_2' = \sqrt{\frac{1}{n}\sum_{j=1}^{s}\frac{(n_j - \hat{n}_j)^2}{n_j}}, \tag{9}$$

gdzie *s* to liczba przedziałów, na które zostały pogrupowane uporządkowane dane, a n_j oraz $\hat{n}_j$ to rzeczywiste oraz szacowane częstości poszczególnych przedziałów.

Dla rozkładu Zengi szacowane częstości wyznaczamy w następujący sposób:

$$\hat{n}_j = \begin{cases} nF\left(x_1 : \hat{\mu}; \hat{\alpha}; \hat{\theta}\right), & j = 1 \\ n\left\{F\left(x_j : \hat{\mu}; \hat{\alpha}; \hat{\theta}\right) - F\left(x_{j-1} : \hat{\mu}; \hat{\alpha}; \hat{\theta}\right)\right\}, & j = 2,3, \dots, s, \end{cases}$$

gdzie $\hat{\mu}, \hat{\alpha}, \hat{\theta}$ są estymowanymi wartościami parametrów funkcji Zengi $f(x : \mu;\ \alpha;\ \theta)$.

Analizowane dane zostały pogrupowane na 20 przedziałów.

4. ZASTOSOWANIE ROZKŁADÓW DAGUMA I ZENGI DO ANALIZY ROZKŁADU DOCHODÓW GOSPODARSTW DOMOWYCH

Rozkłady Daguma i Zengi zostały wykorzystane do analizy dochodów gospodarstw domowych ludności Polski w 2014 r. Rysunek 3 obrazuje różne metody estymacji rozkładu Zengi do danych empirycznych. W tabl. 1 zestawiono miary zgodności dopasowania do danych empirycznych, takie jak wskaźnik podobieństwa struktur oraz indeksy określone wzorami (7), (8), (9) dla rozkładu Daguma i rozkładu Zengi. Tablica 2 zawiera empiryczne oraz teoretyczne charakterystyki rozkładu dochodów mieszkańców Polski.

Do estymacji parametrów rozkładu Zengi użyto metod D'Addaria opartych na dwóch oraz trzech ograniczeniach oraz metody momentów. Wykorzystanie metody największej wiarygodności dla rozkładu Zengi rozważano w pracach Arcagniego (2014), Arcagniego i Porra (2013) oraz Arcagniego i Zengi (2013), jednak otrzymane wyniki były nieznacznie gorsze od wyników otrzymanych prostszymi metodami minimalizującymi różnice między empirycznymi i teoretycznymi charakterystykami rozkładu.

Zebrane wyniki pokazują, że najlepsze dopasowanie do danych empirycznych dla metody D'Addaria z trzema ograniczeniami uzyskuje się, stosując metodę 2:

$$\begin{cases} \hat{\mu} = \bar{x} \\ F(\hat{m} : \bar{x}, \alpha, \theta) = 0{,}5 \\ P(\bar{x}, \alpha, \theta) = \hat{P} \end{cases}.$$

Natomiast stosując metodę D'Addaria z dwoma ograniczeniami, najlepsze dopasowanie otrzymujemy przy użyciu następującej metody:

$$\begin{cases} \hat{\mu} = \bar{x} \\ F(\hat{m} : \bar{x}, \alpha, \theta) = 0{,}5. \end{cases}$$

Metoda momentów daje również zadowalające rezultaty, a wyznaczona analitycznie postać parametrów funkcji gęstości Zengi jest niewątpliwie zaletą stosowania tej metody. Warto zauważyć, że używając metody D'Addaria należy zastosować metody numeryczne.

W wyniku estymacji parametrów metodą największej wiarygodności w pracy Jędrzejczak i Pekasiewicz (2018) wyznaczono parametry rozkładu Daguma: $p = 0{,}747, a = 3{,}12,\ b = 3611{,}02$ dla budżetów gospodarstw domowych ludności Polski w 2014 r.

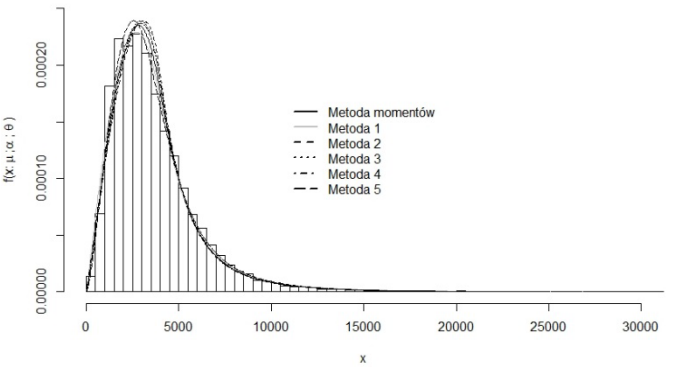
TABLICA 1. WYNIKI ESTYMACJI ROZKŁADÓW DAGUMA I ZENGI

Metoda	Parametr funkcji gęstości			Indeks dopasowania			Współczynnik podobieństwa struktur
	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\theta}$	A_1	A_2	A_2'	w_p
Model Zengi							
Metoda momentów	3755,33	2,2427	3,0303	0,0952	0,1271	0,1375	0,9524
Metody z trzema ograniczeniami: $F(\hat{m}: \bar{x}; \alpha; \theta) = 0{,}5$ $F(\bar{x}: \bar{x}; \alpha; \theta) = \hat{F}(\bar{x})$	3755,33	2,2502	2,9101	0,1112	0,1415	0,1435	0,9444
$F(\hat{m}: \bar{x}; \alpha; \theta) = P$ $P(\bar{x}; \alpha; \theta) = \hat{P}(\bar{x})$	3755,33	2,5693	3,6254	0,0666	0,0905	0,0929	0,9667
$F(\hat{m}: \bar{x}; \alpha; \theta) = 0{,}5$ $Var(\bar{x}; \alpha; \theta) = m_2$	3755,33	2,1202	3,0247	0,0857	0,1338	0,1684	0,9572
$F(\bar{x}: \bar{x}; \alpha; \theta) = \hat{F}(\bar{x})$ $Var(\bar{x}; \alpha; \theta) = m_2$	3755,33	2,1132	2,6991	0,1235	0,1596	0,1684	0,9383
$F(\hat{m}: \bar{x}; \alpha; \theta) = 0{,}5$ $A(\bar{x}; \alpha; \theta) = \hat{A}(\bar{x})$	3755,33	2,1723	2,8501	0,1627	0,2172	0,2779	0,9338
Metody z dwoma ograniczeniami: $F(\hat{m}: \bar{x}; \alpha; \theta) = 0{,}5$	3755,33	2,7943	3,9244	0,0599	0,0815	0,0052	0,9700
$F(\bar{x}: \bar{x}; \alpha; \theta) = \hat{F}(\bar{x})$	3755,33	3,6195	5,0377	0,0733	0,1205	0,1062	0,9633
$P(\bar{x}; \alpha; \theta) = \hat{P}(\bar{x})$	3755,33	2,1469	2,9309	0,0973	0,6023	0,1553	0,9513
$P(\bar{x}; \alpha; \theta) = \hat{P}(\bar{x})$	3755,33	3,1602	1,6269	0,6700	1,0572	0,7787	0,6649
Model Daguma							
Metoda największej wiarygodności	*p*	*a*	*b*	A_1	A_2	A_2'	w_p
	0,747	3,12	3611,02	0,0745	0,0965	0,0974	0,9627

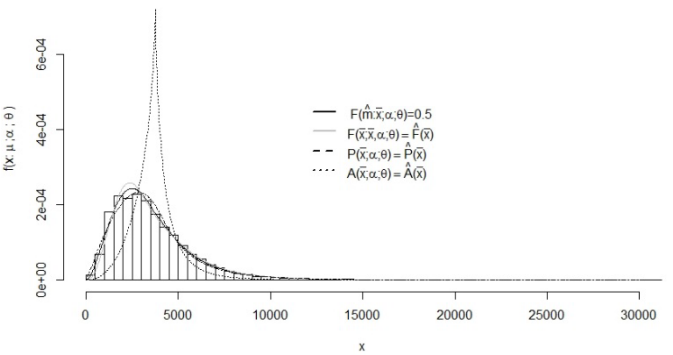
Ź r ó d ł o: opracowanie własne.

Rysunek 3. Aproksymacja modelu Zengi do empirycznego rozkładu dochodów ludności Polski (36929 obserwacji)

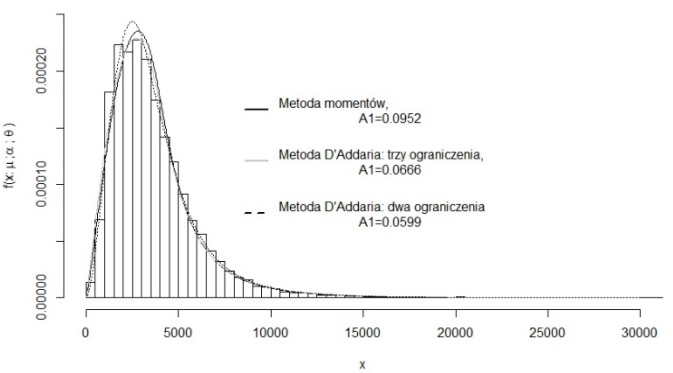
Metoda momentów oraz metoda D'Addaria z trzema ograniczeniami



Metoda D'Addaria z dwoma ograniczeniami

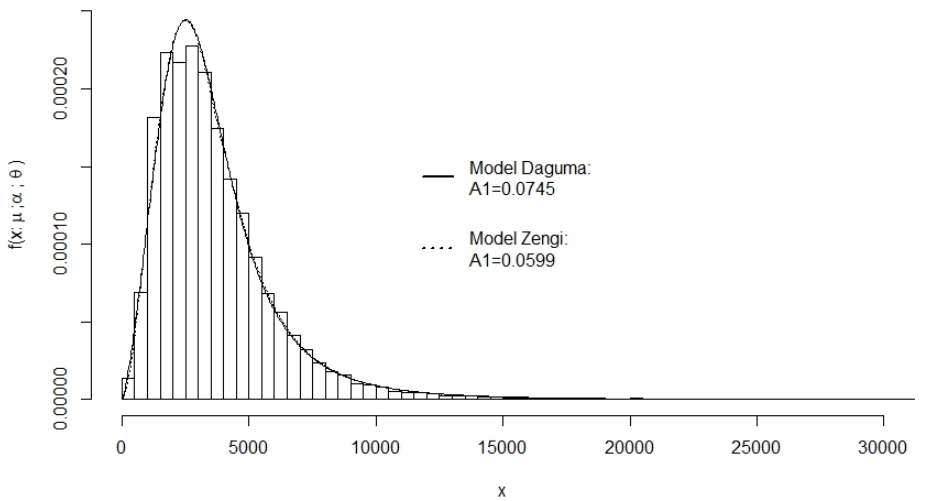


Zestawienie najlepszych metod estymacji



Ź r ó d ł o: opracowanie własne.

Rysunek 4. Aproksymacja rozkładu Daguma oraz rozkładu Zengi do empirycznego rozkładu dochodów ludności Polski (36929 obserwacji)



Ź r ó d ł o: opracowanie własne.

Z przedstawionych rezultatów aproksymacji wynika, że możliwe jest określenie parametrów rozkładu Zengi dla dochodów indywidualnych. O zgodności tego rozkładu z obserwacjami empirycznymi świadczą wysokie wartości współczynników podobieństwa struktur oraz niskie wartości indeksów dopasowania. Podobne rezultaty otrzymał Zenga (2010). Porównując indeksy dopasowania rozkładu dochodów ludności Polski (Jędrzejczak, Trzcińska, 2018) z indeksami dla Włoch, Stanów Zjednoczonych oraz Szwajcarii wyznaczonymi przez Zengę, można zaobserwować, że dla Polski są nieco wyższe. Przyczyn tych rozbieżności należy upatrywać przede wszystkim w różnicach definicji zmiennej dochodowej, która w przytoczonych badaniach jest dochodem ekwiwalentnym, natomiast w badaniu polskim zastosowano dochód rozporządzalny gospodarstwa domowego. Dodatkowy wpływ na wynik mają odmienne schematy losowania próby oraz inne liczebności. Z przeprowadzonej analizy można wywnioskować, że zmiany parametrów α oraz θ rozkładu Zengi objaśniają zmiany wartości rozważonych miar statystycznych, co ilustruje tabl. 2. Parametry α oraz θ są parametrami nierównomierności rozkładu Zengi, określają one postać krzywej koncentracji Lorenza, a tym samym współczynnik Giniego. Trzeci parametr rozkładu Zengi μ jest parametrem skali, więc nie wpływa na wartość współczynnika Giniego (Zenga i in., 2010b).

TABLICA 2. EMPIRYCZNE I TEORETYCZNE CHARAKTERYSTYKI ROZKŁADU DOCHODÓW W POLSCE

Metoda	Współczynnik Giniego	Średnia arytmetyczna	Odchylenie standardowe	Mediana	Kwantyl rzędu 0,1	Kwantyl rzędu 0,95
Model Zengi						
Metoda momentów	0,2704	3755,33	2959,97	3202,09	1319,16	8299,97
$F(\hat{m}\colon \bar{x};\alpha;\theta)=0{,}5$ $F(\bar{x}\colon \bar{x};\alpha;\theta)=\hat{F}(\bar{x})$	0,3399	3755,33	2879,44	3237,46	1359,05	8168,20
$F(\hat{m}\colon \bar{x};\alpha;\theta)=P$ $P(\bar{x};\alpha;\theta)=\hat{P}(\bar{x})$	0,3484	3755,33	2847,59	3158,97	1341,34	8379,67
$F(\hat{m}\colon \bar{x};\alpha;\theta)=0{,}5$ $Var(\bar{x};\alpha;\theta)=m_2$	0,3611	3755,33	3151,08	3158,99	1245,80	8487,26
$F(\bar{x}\colon \bar{x};\alpha;\theta)=\hat{F}(\bar{x})$ $Var(\bar{x};\alpha;\theta)=m_2$	0,3414	3755,33	2959,93	3251,09	1339,54	8159,10
$F(\hat{m}\colon \bar{x};\alpha;\theta)=0{,}5$ $A(\bar{x};\alpha;\theta)=\hat{A}(\bar{x})$	0,3414	3755,33	2959,91	3227,95	1329,58	8224,76
$F(\hat{m}\colon \bar{x};\alpha;\theta)=0{,}5$	0,3458	3755,33	2745,12	3158,99	1378,89	8332,40
$F(\bar{x}\colon \bar{x};\alpha;\theta)=\hat{F}(\bar{x})$	0,3308	3755,33	2510,97	3155,49	1475,78	8207,88
$P(\bar{x};\alpha;\theta)=\hat{P}(\bar{x})$	0,3558	3755,33	3049,53	3195,90	1289,44	8348,12
$A(\bar{x};\alpha;\theta)=\hat{A}(\bar{x})$	0,1300	3755,33	1393,82	3663,77	2356,82	5938,10
Model Daguma						
Metoda największej wiarygodności	0,3445	3775,76	3016,78	3151,42	1364,88	8426,81
Wartości empiryczne	0,3410	3755,33	2959,95	3158,99	1352,47	8150,00

Ź r ó d ł o: opracowanie własne.

Tablica 2 przedstawia oszacowane wartości charakterystyk liczbowych rozkładów teoretycznych oraz ich empiryczne wartości. Jak widać, za pomocą rozkładu Zengi otrzymujemy najlepsze dopasowanie do danych empirycznych (rys. 4). W ramach przeprowadzonych badań rozkład Daguma przeszacowuje wybrane miary, wyjątek stanowi mediana, gdyż w tym przypadku rozkład Daguma nie doszacowuje tej wartości. Rozkład Zengi idealnie szacuje średnią arytmetyczną, wynika to z konstrukcji metod estymacji. Idealną zgodność z danymi empirycznymi dla mediany otrzymujemy dzięki metodzie D'Addaria z dwoma ograniczeniami $\begin{cases}\mu=\bar{x}\\ F(\hat{m}\colon \bar{x};\alpha;\theta)=0{,}5\end{cases}$ oraz trzema ograniczeniami $\begin{cases}\mu=\bar{x}\\ F(\hat{m}\colon \bar{x};\alpha;\theta)=0{,}5\\ Var(\bar{x};\alpha;\theta)=m_2\end{cases}$. Najlepsze oszacowanie odchylenia standardowego otrzymujemy przy zastosowaniu

metody momentów oraz metody D'Addaria z trzema ograniczeniami: $\begin{cases} \mu = \bar{x} \\ F(\bar{x}: \bar{x}; \alpha; \theta) = \hat{F}(\bar{x}); \\ Var(\bar{x}; \alpha; \theta) = m_2 \end{cases}$ $\begin{cases} \mu = \bar{x} \\ F(\hat{m}: \bar{x}; \alpha; \theta) = 0{,}5. \\ A(\bar{x}; \alpha; \theta) = \hat{A}(\bar{x}) \end{cases}$ Najlepsze dopasowanie do danych empirycznych kwantyla rzędu 0,1 otrzymujemy za pomocą metody D'Addaria z trzema ograniczeniami $\begin{cases} \mu = \bar{x} \\ F(\hat{m}: \bar{x}; \alpha; \theta) = 0{,}5 \\ F(\bar{x}: \bar{x}; \alpha; \theta) = \hat{F}(\bar{x}) \end{cases}$, natomiast najlepsze dopasowanie kwantyla rzędu 0,95 otrzymujemy, stosując metodę D'Addaria z trzema ograniczeniami $\begin{cases} \mu = \bar{x} \\ F(\bar{x}: \bar{x}; \alpha; \theta) = \hat{F}(\bar{x}) \\ Var(\bar{x}; \alpha; \theta) = m_2 \end{cases}$. Jeżeli chodzi o oszacowanie współczynnika Giniego, to na najlepsze dopasowanie do danych empirycznych pozwala metoda D'Addaria z trzema ograniczeniami: $\begin{cases} \mu = \bar{x} \\ F(\bar{x}: \bar{x}; \alpha; \theta) = \hat{F}(\bar{x}) \\ Var(\bar{x}; \alpha; \theta) = m_2 \end{cases}$ oraz $\begin{cases} \mu = \bar{x} \\ F(\hat{m}: \bar{x}; \alpha; \theta) = 0{,}5. \\ A(\bar{x}; \alpha; \theta) = \hat{A}(\bar{x}) \end{cases}$ Najgorsze szacowanie wybranych miar otrzymujemy dla metody D'Addaria z dwoma ograniczeniami $\begin{cases} \mu = \bar{x} \\ A(\bar{x}; \alpha; \theta) = \hat{A}(\bar{x}) \end{cases}$, co nie jest zaskoczeniem, gdyż ta metoda estymacji ma najniższą wartość współczynnika podobieństwa struktur oraz najwyższe wartości indeksów dopasowania.

Na podstawie przeprowadzonej analizy można stwierdzić, że metoda zapewniająca najlepszą zgodność dopasowania do danych empirycznych niekoniecznie pokrywa się z metodą najlepszą do szacowania konkretnej charakterystyki. Spostrzeżenie to jest zgodne z wynikami pracy Ostasiewicz (2013). Zauważmy, że stosując rozkład Zengi, w każdym przypadku otrzymujemy lepsze oszacowanie do danych empirycznych wybranych miar niż za pomocą rozkładu Daguma. Potwierdza to zasadność stosowania rozkładu Zengi do opisu dochodów Polaków. Słabą stroną rozkładu Zengi może być jedynie to, że klasyczne metody estymacji parametrów są dość skomplikowane. Chodzi tu m.in. o zastosowanie metody największej wiarygodności, która wiąże się z koniecznością bardzo czasochłonnych obliczeń z użyciem metod numerycznych.

5. WNIOSKI

W artykule przedstawiono aproksymację rozkładu płac i dochodów ludności Polski za pomocą rozkładów Daguma i Zengi. Porównując wykresy funkcji gęstości obu rozkładów, można zauważyć, że funkcja gęstości rozkładu Zengi przyjmuje bardziej różnorodne kształty niż funkcja gęstości rozkładu Daguma. Dzięki temu rozkład Zengi łatwiej dopasowuje się do różnego rodzaju danych

empirycznych. Na podstawie rozważań teoretycznych i przedstawionych wyników estymacji zauważamy, że w odróżnieniu od innych rozkładów trójparametrowych, np. rozkładu Daguma, parametry rozkładu Zengi można interpretować w kategoriach ekonomicznych. Parametr μ jest bezpośrednio interpretowalny jako średni dochód indywidualny lub średni dochód gospodarstwa domowego, natomiast parametry α i θ mogą być interpretowane w kategoriach nierówności dochodowych. W szczególności wartość parametru α jest odwrotnym wskaźnikiem nierówności, który kontroluje kształt rozkładu, podczas gdy θ jest bezpośrednim wskaźnikiem nierówności, który kontroluje rozkład wokół wartości oczekiwanej μ (Arcagni, Porro, 2013).

Do estymacji parametrów rozkładu Daguma zastosowano metodę największej wiarygodności, a dla rozkładu Zengi – metodę momentów i metodę D'Addaria z dwoma i trzema ograniczeniami. Z analizy metod estymacji przeprowadzonych dla parametrów funkcji gęstości Zengi wynika, że najlepsze dopasowanie uzyskuje się za pomocą metody D'Addaria z dwoma ograniczeniami: $\hat{\mu} = \bar{x}$ oraz $F(\hat{m}; \bar{x}, \alpha, \theta) = 0{,}5$. Metoda momentów przynosi również zadowalające efekty. Na podstawie przeprowadzonej analizy można stwierdzić, że metoda zapewniająca najlepszą zgodność dopasowania do danych empirycznych niekoniecznie pokrywa się z metodą najlepszą do szacowania konkretnej charakterystyki. Spostrzeżenie to jest zgodne z wynikami pracy Ostasiewicz (2013).

Z przedstawionych rezultatów wynika, że możliwe jest efektywne dopasowanie rozkładu dochodów z użyciem modelu Zengi. Ten wniosek jest zgodny z wynikami podobnych badań prowadzonych nad rozkładami dochodów przez Zengę (2010) oraz Zengę i in. (2010a, 2010b, 2010c).

Porównanie rozkładów Daguma i Zengi pokazuje, że rozkład Zengi lepiej opisuje rozkład dochodów ludności Polski. Rozkład Zengi spełnia kryteria stawiane teoretycznym rozkładom dochodów. Przejrzysta interpretacja ekonomiczna parametrów rozkładu Zengi jest dodatkowym argumentem skłaniającym do dalszych badań dotyczących przydatności tego rozkładu do opisu rozkładu płac i dochodów ludności Polski.

LITERATURA

Arcagni A., (2014), Zenga Distribution: Parameters Estimation with Constraints on Synthetic Inequality Measures, *Statistica & Applicazioni*, 12(1), 17–39.

Arcagni A., Porro F., (2013), On the Parameters of Zenga Distribution, *Statistical Methods and Applications*, 22(3), 285–303. DOI: 10.1007/s10260-012-0219-y.

Arcagni A., Zenga M., (2013), Application of Zenga's Distribution to a Panel Survey on Household Income of European Member States, *Statistica & Applicazioni*, 2(1), 79–102. DOI: 10.1400/209741.

Bandurian R., McDonald J. B., Turley R. S., (2003), A comparison of parametric models of income distribution across countries and over time, *Estadistica*, 55, 135–152.

Brzeziński M. (2013), Parametric modelling of income distribution in Central and Eastern Europe, *Central European Journal of Economic Modelling and Econometrics*, 35, 207–230.

Dagum C., (1977), A New Model of Personal Income Distribution: Specification and Estimation, *Economie Appliquée*, 30, 413–437.

Dagum C., (1990), Generation and Properties of Income Distribution Functions, *Proceedings of the Second International Conference on Income Distribution by Size: Generation, Distribution, Measurement and Applications*.

Dastrup S. R., Hartshorn R., McDonald J. B, (2007), The Impact of Taxes and Transfer Payments on the Distribution of Income: A Parametric Comparison, *Journal of Economic Inequality*, 5, 353–369. DOI: 10.1007/s10888-006-9039-3.

D'Addario R., (1934), *Sulla Misura Della Concentrazione dei Redditi*, Poligrafico dello stato, Roma.

D'Addario R., (1939), Un Metodo Per la Rappresentazione Analitica Delle Distribuzioni Statistiche, *Ammali dell 'Instituto di Statistica dell Universita di Bari*, 16, 3–56.

Domański C., (1973), Struktura płac pracowników w województwach grodzkich według działów gospodarki narodowej, *Studia Prawno-Ekonomiczne*, 11, 149–163.

Domański C., Jędrzejczak A., (1998), Maximum Likelihood Estimation of the Dagum Model Parameters, *International Advances in Economic Research*, 4, 243–252.

Domański C., Jędrzejczak A., (2005), *Analiza rozkładów płac i dochodów w Polsce w przekroju terytorialnym*, Wydawnictwo Uniwersytetu Łódzkiego, Łodź.

Edgeworth F. Y., (1898), On the Representation of Statistics by Mathematical Formulae, *Journal of the Royal Statistical Society*, 61, 670–700.

Jagielski M., Kutner R., (2010), Study of Households' Income in Poland by Using the Statistical Physics Approach, *Acta Physica Polonica A*, 117(4), 615–618. DOI: 10.12693/APhysPolA. 117.615.

Jędrzejczak, A., (1990), Uwagi o zastosowaniu rozkładu Daguma do badania rozkładów płac, *Wiadomości Statystyczne*, 7, 23–25.

Jędrzejczak A. (1999, maszynopis), *Statystyczna analiza nierównomierności rozkładów płac i dochodów* [praca doktorska, Uniwersytet Łódzki].

Jędrzejczak A., Pekasiewicz D., (2018), Analysis of the Properties of Selected Inequality Measures Based on Quantiles with the Application to the Polish Income Data, w: M. Papież, S. Śmiech (eds.), *The 11th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Conference Proceedings*, 113–122.

Jędrzejczak A., Pekasiewicz D., Zieliński W., (2018), Comparison of Estimation Methods for the Dagum Distribution Parameters, w: M. Papież, S. Śmiech (eds.), *The 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Conference Proceedings,* 180–189.

Jędrzejczak A., Pekasiewicz D., Zieliński W., (2019), Confidence Interval for Quantile Ratio of the Dagum Distribution, *Revstat – Statistical Journal*.

Jędrzejczak A., Trzcińska K., (2018), Application of the Zenga distribution to the analysis of household income in Poland by socio-economic group, *Statistica & Applicazioni*, 16(2), 123–140.

Kleiber C., Kotz S., (2003), *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley, Hoboken.

Kordos J., (1968), *Metody matematyczne badania I analizy rozkładu dochodów ludności*, Główny Urząd Statystyczny, Warszawa.

Kordos J., (1973), *Metody analizy i prognozowania rozkładów płac i dochodów ludności*, Główny Urząd Statystyczny, Warszawa.

Kordos J., (1990), Research on Income Distribution by Size in Poland, w: C. Dagum, M. Zenga (eds.), *Income and Wealth Distribution, Inequality and Poverty*, Springer, New York, Berlin, London and Tokyo, 335–351.

Kot S. M., (1999), *Analiza ekonometryczna kształtowania się płac w Polsce w okresie transformacji*, Wydawnictwo Naukowe PWN, Warszawa–Kraków.

Kot S. M., (2000), Ekonometryczne modele dobrobytu, Wydawnictwo Naukowe PWN, Warszawa–Kraków.

Lange O., (1967), *Wstęp do ekonometrii*, Państwowe Wydawnictwo Naukowe, Warszawa.

Łukasiewicz P., Orłowski A., (2004), Probabilistic Models of Income Distributions, *Physica A*, 344, 146–151. DOI: 10.1016/j.physa.2004.06.106.

March L., (1898), Quelques Exemples de Distribution de Salaries, *Journal de la Société Statistique de Paris*, 39, 193–206, 241–248.

Ostasiewicz K. (2013), Adekwatność wybranych rozkładów teoretycznych dochodów w zależności od metody aproksymacji, *Przegląd Statystyczny*, 60(4), 499–521.

Pareto V., (1897), *Cours d'economie Politique,* Rouge, Lausanne–Paris.

Polisicchio M., (2008), The Continuous Random Variable with Uniform Point Inequality Measure I(p), *Statistica & Applicazioni*, 6(2), 137–151.

Salamaga M., (2016), Badanie wpływu metody estymacji teoretycznych modeli rozkładu dochodów na jakość aproksymacji rozkładu dochodów mieszkańców Krakowa, *Zeszyty Naukowe. Uniwersytet Ekonomiczny w Krakowie*, 3(951), 63–79. DOI: 10.15678/ZNUEK.2016.0951.0305.

Vielrose E., (1960), *Rozkład dochodów według wielkości*, Polskie Wydawnictwo Gospodarcze, Warszawa.

Wąsik B., (1967), Dwu- i trójparametryczny rozkład logarytmiczno-normalny jako aproksymacja rozkładów zarobków pracowników gospodarki uspołecznionej 1955–1965, *Przegląd Statystyczny*, 14(4), 409–424.

Wiśniewski J., (1934), *Rozkład dochodów według wysokości*, Instytut Badań Koniunktur Gospodarczych i Cen, Warszawa.

Zenga M. M., (2007), Inequality Curve and Inequality Index Based on the Ratios Between Lower and Upper Arithmetic Means, *Statistica & Applicazioni*, 5(1), 3–28.

Zenga M. M., (2010), Mixture of Polisicchios Truncated Pareto Distributions with Beta Weights, *Statistica & Applicazioni*, 8(1), 3–25.

Zenga M. M., Pasquazzi L., Zenga Ma. (2010a), First Applications of a New Three Parameter Distribution for Non-Negative Variables, *Rapporto di Ricerca* N. 188, Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali – Universita degli Studi di Milano-Bicocca.

Zenga M. M., Polisicchio M., Zenga Ma., Pasquazzi L., (2010b), More on a New Three-Parameter Distribution for Non-Negative Variables, *Rapporto di Ricerca* N. 187, Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali – Universita degli Studi di Milano-Bicocca.

Zenga M. M., Porro F., Arcagni A., (2010c), Method of Moments for Zenga Distribution, *Rapporto di Ricerca* N. 193, Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali – Universita degli Studi di Milano-Bicocca.

DOI: 10.5604/01.3001.0014.0953

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 4 – 2019

Krzysztof DMYTRÓW[1]

Calibration of attributes influence in the process of real estate mass appraisal by using decision-making methods[2]

Abstract. *There are situations in the real estate market in which a large number of properties have to be valued at the same time. In such cases it is advisable to use mass valuation methods. These methods involve estimating the value of a property on the basis of the values of the attributes defining it. The aim of the paper is to calibrate the influence of attributes on unit values of properties in mass appraisal in order to minimise the valuation error. The research was conducted for 318 residential properties located in Szczecin. The Szczecin Algorithm of Real Estate Mass Appraisal was used along with the econometric, statistical and expert approaches. The econometric approach is based on the ridge regression model, the statistical approach on the partial Kendall τ correlation coefficients, and the expert approach on the AHP method. The quadratic programming was co-employed with the statistical and expert approaches in order to minimise the mean square error* (MSE) *of the valuations. The econometric and statistical approaches with the minimisation of the* MSE *generated best results. The least accurate results were obtained by means of the statistical and expert approaches without the minimisation of the* MSE*. However, even though the optimisation of the* MSE *improves the quality of valuations, it also narrows down their volatility, which might make the valuation of properties from the outside of a given database more problematic.*

Keywords: real estate mass appraisal, real estate attributes, AHP method, statistical and econometric methods of real estate mass appraisal, quadratic programming.

JEL Classification: C34, C44, R30

Kalibracja wpływu atrybutów w procesie wyceny masowej nieruchomości za pomocą metod decyzyjnych

Streszczenie. *Zdarzają się sytuacje, gdy jednocześnie należy wycenić dużą liczbę nieruchomości. W takich przypadkach wskazane jest stosowanie metod masowej wyceny nieruchomości. Za ich pomocą wycenia się wartości nieruchomości na podstawie wartości definiujących je atrybutów. Celem artykułu jest taka kalibracja wpływu atrybutów na*

[1] University of Szczecin, Institute of Economics and Finance, ul. Mickiewicza 64, 71-101 Szczecin, e-mail: krzysztof.dmytrow@usz.edu.pl, ORCID: https://orcid.org/0000-0001-7657-6063.

[2] The research was conducted within the project financed by the National Science Centre, Project No. 2017/25/B/HS4/01813.

wartości jednostkowe nieruchomości w masowej wycenie, aby uzyskać jedynie minimalne błędy wycen. Badanie przeprowadzono dla 318 nieruchomości mieszkaniowych zlokalizowanych w Szczecinie. Zastosowano Szczeciński Algorytm Masowej Wyceny Nieruchomości wraz z podejściem ekonometrycznym, statystycznym i eksperckim. Podejście ekonometryczne oparto na modelu regresji grzbietowej, podejście statystyczne na cząstkowych współczynnikach korelacji τ Kendalla, natomiast podejście eksperckie na metodzie AHP. W celu zminimalizowania średniokwadratowego błędu wycen (MSE), zastosowano programowanie kwadratowe wraz z podejściem statystycznym i eksperckim. Najlepsze wyceny osiągnięto przy zastosowaniu podejścia ekonometrycznego i statystycznego z minimalizacją MSE. Najmniej dokładne wyniki uzyskano przy zastosowaniu podejścia statystycznego i eksperckiego bez minimalizacji MSE. Z jednej strony optymalizacja MSE poprawia jakość wycen, ale z drugiej zawęża ich zmienność, co może sprawić, że wycena nieruchomości dla innych danych może być problematyczna.

1. INTRODUCTION

Real estate appraisal can be performed individually, i.e. separately for each property, or by means of mass appraisal, i.e. when many properties are valued simultaneously. Each of the two above-mentioned methods has its advantages and disadvantages. The main advantage of the individual appraisal is that just one property or a limited number of properties are appraised by a real estate appraiser directly, which allows taking into account all the specific features that are sometimes unique and can be attributed only to one given property, therefore the appraisal can be very precise. On the other hand, however, the process of such valuation is time-consuming and the number of real estates that can be appraised individually in a defined period of time is limited. Individual appraisals are the most frequently performed procedures that use applicable valuation rules resulting from the law, a number of professional standards, basic and specialist valuation standards and interpretative notes (Źróbek and Bełej, 2000).

On the other hand, we can talk about real estate mass appraisal when the following conditions are met simultaneously (Hozer et al., 2002):

- the object of valuation is a large number of properties of the same type;
- the valuation is carried out using the same method for each property, which yields comparable results;
- all properties are valued simultaneously.

However, it should be noted that the two above-mentioned types of appraisal (individual and mass) are not considered as alternative or substitutional, but complementary. The individual real estate appraisal is usually carried out when there is a need for determining the value of a specific property, for the purpose of a purchase or sale of a property, for insurance purposes or to assess losses caused by random incidents, robberies, etc. The real estate mass appraisal, on the other hand, is adopted for the purposes of (Hozer et al., 1999):

- carrying out the revaluation of annual fees for the perpetual usufruct,
- estimation of the economic effects of adopting or amending local spatial development plans,
- monitoring the value of real estate that secures a bank's credit exposures in order to calculate LtV for the bank's credit portfolio,
- universal property taxation,
- other necessities, such as the sale of residential property from municipal resources, expropriation for linear investments, etc.

In order to perform mass real estate appraisal, all properties (of the same type) have to be definable by the same attributes. Therefore, it is very important to prepare a good, reliable database prior to the valuation process. As in the case of individual property appraisal, mass appraisal also requires the work of property appraisers, who have to determine reliable values of attributes. Real estate mass valuation should be carried out by means of quantitative methods. Literature distinguishes four main groups of quantitative methods used in real estate mass appraisal (Kauko and D'amato, 2008):

- model-driven methods;
- data-driven methods;
- methods based on machine learning;
- expert methods.

Model-driven methods constitute a set of classical, quantitative methods such as standard regression models, hedonic regression models or spatial regression models. Data-driven methods include non-parametric Geographically Weighed Regression (GWR) models. They are closely connected with machine learning techniques, such as the Artificial Neural Network, fuzzy sets, genetic algorithms, ridge regression, lasso regression, random forests, regression trees, etc. Nowadays, with the development of programming languages and increasing computing capability of computers, these methods are becoming increasingly popular (Ćetković et al., 2018). The first three groups of the above-mentioned quantitative methods require access to a large amount of data. In the case of incomplete databases (where some possible values of attributes are missing), it is either impossible or very difficult to obtain a reliable model. In such cases, it is necessary to consult experts about the influence of attributes on the value of a property. This can be done by means of the expert approach, based on multiple-criteria decision making (MCDM) methods. It would be hard to list all of the MCDM methods, because of their large number (Saaty and Ergu, 2015). They can be divided into two main groups (Hwang and Yoon, 1981):

- Multiple Attribute Decision Making (MADM);
- Multiple Objective Decision Making (MODM).

In the first group, the criteria are defined by the attributes, while in the second group – by the objectives. Also, in the first group, all the alternatives are explicitly known, while in the second, not necessarily (there could be an infinite number of them). Problems that arise while employing methods from the second group can be solved by means of the multiple objective mathematical programming procedures. As regards the methods from the first group, the decision-maker is usually interested in sorting, ranking or classifying alternatives. Due to the fact that the criteria are defined by attributes here, these methods are useful for the expert approach of specifying the influence of attributes on the value of properties. Methods used for solving problems within this group are called discrete MCDM methods or MADM methods. The best known of them are: AHP (Analytical Hierarchy Process), ANP (Analytic Network Process), REMBRANDT (Ratio Estimation in Magnitudes or deciBells to Rate Alternatives which are Non-DominaTed), DEMATEL (DEcision MAking Trial and Evaluation Laboratory), ELECTRE (ELimination Et Choix Traduisant la REalité), PROMETHEE (Preference Ranking Organization METHod for Enrichment of Evaluations), COPRAS (Complex Proportional Assessment of Alternatives), MAUT (Multi-attribute Utility Theory), MAVT (Multiattribute Value Theory), SAW (Simple Additive Weighting), VIKOR, TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) and many others (Saaty and Ergu, 2015). Also, there are MADM methods that originate from the multivariate statistical analysis, based on Hellwig's composite measure of development (Nermend, 2017).

If we want to employ experts to assess the influence of attributes on the value of real estate, it could be done by means of weights. In the expert approach, weights could be assessed directly by experts, but this would be highly subjective. Another method of obtaining weights of criteria (attributes) on the value of real estate is applying an appropriate multiple-criteria decision making technique. The question which technique to choose for this purpose could be answered by studying the problem closey. There are many publications which provide guidelines on how to choose the best-fitting decision-making technique. Guitouni and Martel (1998) presented the characteristics of almost 30 MCDM methods along with the recommendations as to which of them to use in which situation. Saaty and Ergu (2015) presented 16 criteria by which various MCDM methods were evaluated and compared. A comprehensive comparison of many MCDM methods was also carried out by Evangelos and Triantaphyllou (2000). If the aim of a study is to obtain the vector of weights of the criteria, the AHP method would be a good choice (Guitouni and Martel, 1998, p. 508; Trzaskalik, 2014, p. 274-275). The AHP is also a method most widely used in the real estate market of all the MCDM methods. It was applied, for example, to forecasting values of properties (Yalpir, 2014), establishing the weights of real estate attributes (Kozioł-Kaczorek, 2012), the appraisal of properties for purchase purposes (Ball and Srnivasan, 1994) and to making purchase-related decisions (Saaty, 1990). The main advantage of the expert methods over the other above-mentioned ones is the fact that they can be used even if data is incomplete.

However, they also have several disadvantages. For example, instead of the real relationship between the property's attributes and its value, they reflect subjective (although supported by experience) estimation of this relationship by an expert. Such approach might be satisfactory in some cases, but generally it is not as good as when we can assess the above-mentioned relationship on the basis of full data. Also, estimations done by different experts tend to vary. It is assumed that real estate appraisers should assess the states and influence of attributes on the value of real estate in a similar way (this assumption is based on another assumption, namely that if the same real estate is evaluated by several appraisers, the obtained values should be approximately the same), but it is not always true. Therefore, expert methods should be regarded as a complement to other methods of mass real estate appraisal, rather than as an alternative.

The aim of the paper is to calibrate the influence of attributes on the unit value of properties in mass appraisal in order to minimise the valuation error. In the study presented below, properties were appraised by means of the Szczecin Algorithm of Real Estate Mass Appraisal (SAREMA) with the application of the econometric, statistical and expert (based on the AHP method) approaches. The quadratic programming was used in order to minimise the mean square error of valuations in the statistical and expert approaches.

2. RESEARCH METHODOLOGY

The following approaches of the calibration of influence of attributes were adopted in the research:

- econometric approach;
- statistical approach;
- AHP method;
- quadratic programming.

The starting point of each approach is the Szczecin Algorithm of Real Estate Mass Appraisal (SAREMA) (Hozer et al., 1999). In this algorithm, the unit value (or the transactional price) of a property is appraised by means of the following formula:

$$\hat{v}_{ji} = wwr_j \cdot v_b \cdot \prod_{k=1}^{K} \prod_{p=1}^{k_p} (1 + a_{kpi}), \tag{1}$$

where:

$\hat{v}_{ji}$ – unit market (or cadastral) value (value of 1 m^2) of the i-th property ($i = 1, 2, \ldots, n$) in the j-th location attractiveness zone ($j = 1, 2, \ldots, J$),

v_b – basic value of 1 m^2,
a_{kpi} – impact of the p-th state of the k-th attribute on the i-th property,
wwr_j – market value ratio in the i-th location attractiveness zone,
n – number of properties,
J – number of attractiveness zones,
K – number of attributes,
k_p – number of states of the k-th attribute.

The basic value (v_b) can be estimated in several ways. It can be the theoretical value of 1 m^2 of a property with the worst states of attributes in the cheapest location attractiveness zone. It can also be the value of 1 m^2 of the cheapest property in the appraised area. However, this approach is not suitable if we want to extend the algorithm to real estate from beyond the database currently being analysed. The reason is that it is strictly connected with the market value ratio (wwr_j). This measure informs us about the influence of the widely understood location on the value of a property. The location attractiveness zones are spatial units where properties of the same purpose with the same values of attributes should obtain similar values. In order to estimate the market value ratios, we need to have a set of representative properties that have to be appraised by real estate appraisers. Next, the values of the same properties have to be calculated using the algorithm. It is done by the following formula:

$$\hat{v}_{hji} = v_b \cdot \prod_{k=1}^{K} \prod_{p=1}^{k_p} (1 + a_{kpi}), \tag{2}$$

where $\hat{v}_{hji}$ – unit value of the i-th representative property in the j-th location attractiveness zone.

The value of wwr_{ji} for each representative property is calculated in the following way:

$$wwr_{ji} = \frac{v_{rji}}{\hat{v}_{hji}}, \tag{3}$$

where v_{ri} – unit value of the i-th representative property in the j-th location attractiveness zone.

Finally, market value ratios for each location attractiveness zone are calculated using the following formula:

$$wwr_j = \sqrt[l_j]{\prod_{i=1}^{l_j} wwr_{ji}}, \tag{4}$$

where l_j is the number of properties in the j-th attractiveness zone. Because the market value ratio is understood as the so-called 'location premium', its value for each location attractiveness zone should be no less than 1. The algorithm has a multiplicative form, therefore the geometric mean is used.

The most important and difficult part of applying the algorithm described by Formula (1) is the estimation of the impact of the p-th state of the k–th attribute on the i-th property (a_{kpi}). It can be done using various approaches. The econometric approach, being the modification of Formula (1), was proposed by Doszyń (2018). In this approach, logarithms of both sides of Equation (1) were calculated, and the error term was added:

$$ln(v_{ji}) = \alpha_0 + \sum_{k=1}^{K}\sum_{p=2}^{k_p} \alpha_{kp} x_{kpi} + \sum_{j=2}^{J} \alpha_j te_j + u_i, \tag{5}$$

where:

α_0 – intercept parameter (logarithm of the basic unit value),
α_{kp} – impact of the p-th state of attribute k,
x_{kpi} – dummy variable for the p-th state of attribute k for the i-th property ($i = 1, 2, \dots, n$),
α_j – (the logarithm of) the market value coefficient for the j-th location attractiveness zone,
te_j – dummy variable equal to 1 for the j-th location attractiveness zone (and zero for others),
u_i – error term.

By means of Formula (5), we represent the SAREMA in the econometric form. Real estate attributes are measured on the ordinal scale, therefore dummy variables were used. Variables determining the location attractiveness zones are categorical, therefore they are also represented as dummy variables. The main advantage of applying the econometric model is that the influence of each state of every attribute and location attractiveness zone on the unit value is analysed separately and independently. The main disadvantage of this approach is that there are many explanatory variables, which may cause problems with collinearity.

The next approach used in the study is the statistical one. It is based on the relationship between the real estate attributes and the value of 1 m^2 of a piece of real estate, for the representative properties. Because real estate attributes are measured on the ordinal scale, the rank correlation coefficient was applied. The two rank correlation coefficients that are most widely used are the Spearman and the Kendall coefficients. It is possible to use both of them; however, the Spearman coefficient calculates the differences between ranks, which is methodologically incorrect for the ordinal scale. Therefore, the Kendall rank coefficient is a better choice for this kind of data (Doszyń, 2017; Foryś and Gaca, 2016). Due to the fact that the number of cases is higher than the number of variants for each attribute, the Kendall rank coefficients with tied ranks (τ_B) was used (Parker et al., 2011):

Since the attributes may be correlated, partial τ_B Kendall coefficients were calculated:

$$\tau_{yx.z} = -\frac{R_{\mathrm{yx}}}{\sqrt{R_{\mathrm{yy}} \cdot R_{\mathrm{xx}}}}, \tag{6}$$

where:

R_{yx} – determinant of a matrix cofactor obtained by removing the row corresponding to the explained variable y and the column corresponding to the explanatory variable x,
R_{yy} – determinant of a matrix cofactor obtained by removing the row and column corresponding to the explained variable y,
R_{xx} – determinant of a matrix cofactor obtained by removing the row and column corresponding to the explanatory variable x.

Having calculated partial correlation coefficients, we can proceed to the estimation of weights of each attribute in the algorithm (Kolenda, 2006):

$$w_k = \frac{\left|\tau_{yx_k.z}\right|}{\sum_{k=1}^{K}\left|\tau_{yx_k.z}\right|} \cdot 100\%, \tag{7}$$

where:

k – number of the analysed attribute,
K – number of attributes.

The expert approach is based on the AHP method. It was developed by Thomas L. Saaty (1980). It involves pairwise comparisons of the decision criteria, which in this study were the attributes of properties. Pairwise comparisons between the attributes were performed on the basis of a survey, addressed to

four real estate appraisers. Instead of applying the 9-point Saaty scale (Brunelli, 2015), the appraisers suggested using a shorter, 4-point scale, which measured the dominance of attribute 1 over attribute 2:

- attribute 1 has a definitely greater impact (4);
- attribute 1 has a noticeably greater impact (3);
- attribute 1 has a slightly greater impact (2);
- the attributes are indifferent (1);
- attribute 2 has a slightly greater impact (1/2);
- attribute 2 has a noticeably greater impact (1/3);
- attribute 2 has a definitely greater impact (1/4).

The results of the pairwise comparisons c_{kl}, where k, l – compared attributes, $(k, l = 1, 2, \ldots, K; k \neq l)$ between the attributes are placed in the AHP matrix.

If the obtained AHP matrix is consistent, then the weights of attributes can be estimated. The weights obtained in the statistical and expert approaches are used to assess the impact of the p-th state of the k-th attribute $(1 + a_{kp})$ in Formulas (1) and (2):

$$1 + a_{kp} = e^{\ln\left(\frac{v_{\max}}{v_b}\right) \cdot u_{kp}} = \left(\frac{v_{\max}}{v_b}\right)^{u_{kp}} \tag{8}$$

where:

v_{max} – maximum, the theoretical value of 1 m^2 of a property with the best states of attributes in the most expensive location attractiveness zone,

u_{kp} – influence of the weight of the p-th state of the k-th attribute, calculated in the following way:

$$u_{kp} = \frac{w_k}{k_p - 1}(p - 1). \tag{9}$$

Formula (15) is the author's approach to the estimation of the impact of attributes on a unit value of real estate. The ratio $\frac{v_{\max}}{v_b}$ is used in order to transfer the range of unit values of properties onto the valuated ones. The main assumption of Formulas (8) and (9) is that firstly, for the poorest state of a property's attribute ($p = 1$), the value of $1 + a_{kp} = 1$, and secondly, the higher state of the attribute, the higher the value of Formula (8). Formula (9) assumes that the transitions between the states of attributes are (relatively) linear.

The main advantage of the statistical approach is that it reflects the true relationships between the attributes and the unit value of a property. Its main drawback, however, is that in the first stage, the impacts of the p-th state of the k-th attribute $(1 + a_{kp})$ are estimated, while the market value ratios (wwr_j) are calcu-

lated in the second stage. It might lead to a situation where values $1 + a_{kp}$ would depend not only on the relationships between the attributes and the unit value, but also on location. The influence of location is considered in the second stage. Therefore the influence of location on the unit value of a property might be biased. The expert approach should be free from this disadvantage, but the assessment of the relationship between the attributes and the unit value of a property in this approach is subjective. The main advantage of the expert approach is that it can be used regardless of the availability of data.

The statistical and expert approaches are the bases for the application of the quadratic programming to the minimisation of the valuation errors. It is assumed that the highest value of a property is obtained for the best states of attributes, and the lowest value of a property – for the poorest states of attributes. For a perfect database (where properties appraised by experts have values corresponding to the states of attributes), the valuation error should have a minimal value without the need of optimisation (because the algorithm always satisfies the assumption that better states of attributes generate higher values of properties). However, databases (especially from the real estate market) are usually far from being perfect. Therefore, the ratio $\frac{v_{\max}}{v_{\mathrm{b}}}$, obtained by the values set by appraisers, does not necessary reflect the real variability of the values of properties. This ratio, then, will be the subject of optimisation in quadratic programming. The values of the representative properties were appraised both by the appraisers and by the algorithm. The quadratic programming was used to minimise the Mean Square Error (MSE):

$$1 + a_{kp} = e^{\ln\left(\frac{v_{\max}}{v_b}\right)\cdot u_{kp}} = \left(\frac{v_{\max}}{v_b}\right)^{u_{kp}} \tag{10}$$

where:

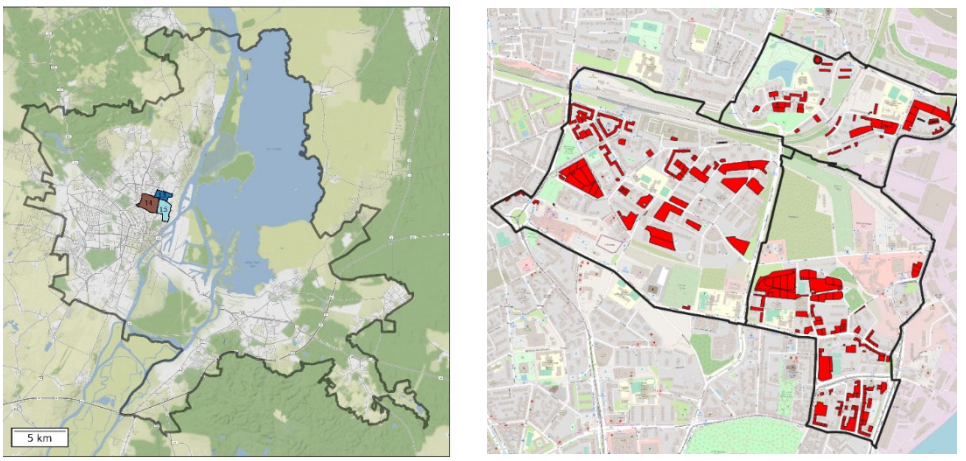
v_{ji} – real unit value of the property determined by the appraiser,
$\hat{v}_{ji}$ – theoretical unit value of the property determined by the algorithm (1).

The quadratic programming was not used in the econometric approach, because the procedure of estimation of the econometric model assumes the minimisation of the sum of squares, which is virtually the same as the *MSE*.

3. DATA USED IN THE STUDY

The research was based on 318 properties located in Szczecin. All of them served housing purposes and were located in three location attractiveness zones (numbered 13, 14 and 15). Their spatial distribution is presented in Figure 1.

Figure 1. Spatial distribution of valued properties



S o u r c e: author's work based on data from the Central Office of Geodesy and Cartography.

All properties were described by five attributes:

- area: 1 – small (up to 500 m^2), 2 – average (500 – 1200 m^2), 3 – large (over 1200 m^2);
- access to utilities: 1 – none, 2 – partial, 3 – full;
- accessibility of public transport: 1 – poor, 2 – average, 3 – good;
- quality of surroundings: 1 – onerous, 2 – unfavourable, 3 – average, 4 – favourable;
- attractiveness of the shape of a plot: 1 – low, 2 – average, 3 – high.

The interpretation of the first four attributes seems relatively straightforward, thus not requiring further explanation. However, the last attribute – 'the attractiveness of the shape of a plot' needs an explanation. It is assumed that that the optimal shape of a plot is a rectangle with the side length ratio of 3:2 (Dmytrów et al., 2018). All the plots were characterised by their area and circumference, among other features. For the circumference, the hypothetical area assuming the optimal rectangle was calculated and compared with the real area. When the ratio of the real area and the hypothetical one was higher than 0.9, the value of the attribute was 3. When it was between 0.5 and 0.9, the value of the attribute was 2, and when it was lower than 0.5, the value of the attribute was 1.

All the 318 properties had full access to utilities, therefore this attribute was removed from the econometric and statistical approaches, as it was impossible to measure its impact on the unit value. It was retained for the expert approach, though, because this approach allows measuring the influence of access to utilities on the value of a property.

Out of the set of 318 properties, 30 representative ones were selected. It should be explained here that the term 'representative property' is not understood in the sense of the representative method. Representative properties are selected in order to ensure that the whole range of attributes in every location attractiveness zone are taken into account in the study. In the analysis presented in this paper it was done by means of stratified sampling from sets of properties with the same values of attributes and in the same attractiveness zones. The borders of location attractiveness zones were determined by experts. The assumption behind this process is that the location attractiveness zones should consist of similar properties, i.e. for the given values of attributes, the values of properties should be roughly the same. But, one might ask, if the goal was to appraise 318 properties (a relatively small population), what was the purpose of selecting 30 of them? It is because of the method – the researcher who uses the SAREMA, has at his/her disposal only these representative properties on the basis of which the market value ratios are estimated. The number of representative properties is usually small, because individual appraisal is time-consuming and expensive. In this study, the representative properties were valued by appraisers and used to estimate the market value ratios in the three analysed location attractiveness zones, by means of equations (2) – (4) (in the statistical and expert approaches). In the statistical approach, they were also used to estimate the values of the partial Kendall τ_B correlation coefficients and the impact of states of attributes on the unit value of a property. In the econometric approach, these representative properties were used to estimate the parameters of the model (5).

The results obtained for the representative properties were then used to appraise all the 318 properties. In order to verify the effectiveness of the SAREMA, all the 318 properties were also valued by real estate appraisers, therefore it was possible to calculate valuation errors for the whole population. The basic error measure was the percentage error (PE):

$$PE_{ji} = \frac{v_{ji} - \hat{v}_{ji}}{v_{ji}} \cdot 100\%. \tag{11}$$

The optimal value of the percentage error is 0%. For all properties, the mean percentage error (MPE) was calculated:

$$MPE = \frac{\sum_{i=1}^{n} PE_{ji}}{n}. \tag{12}$$

The mean percentage error shows if valuations are biased. Its optimum value is 0%. If it is positive, it means that the valuations are on the average underestimated, and if it is negative, it means they are overestimated.

The relative accuracy of valuations was measured using the mean absolute percentage error ($MAPE$):

$$MAPE = \frac{\sum_{i=1}^{n}\left|PE_{ji}\right|}{n}. \tag{13}$$

The closer the value of $MAPE$ to 0%, the better – it means that valuations are closer to the real values of properties. The value of $MAPE$ indicates the mean percentage deviation of values of properties obtained by means of the algorithm from their real unit values.

The absolute accuracy of valuations was measured by means of the root mean square error ($RMSE$):

$$RMSE = \sqrt{MSE}. \tag{14}$$

The smaller the value of $RMSE$, the better. It carries information about the mean absolute deviation of the values of properties obtained by means of the algorithm from their real unit values.

The final measure of accuracy of the valuations were the shares of properties for which the percentage errors (PE) fell within the range of ±5%, ±10% and ±15%.

4. EMPIRICAL RESULTS

In the first step of the analysis, 30 representative properties were valued. The following approaches were adopted:

- econometric (Ec);
- statistical (St);
- expert (Ex);
- statistical with the minimisation of the MSE (St_{MSE});
- expert with the minimisation of the MSE (Ex_{MSE}).

The basic value of 1 m^2 (v_b) was estimated by property appraisers at 279 PLN. Theoretical maximum value of 1 m^2 (v_{max}) was estimated at 708 PLN. Therefore the $\frac{v_{\max}}{v_{\mathrm{b}}}$ ratio equalled 2.54; it means that the most expensive property that had the best states of attributes and was located in the most expensive location attractiveness zone was expected to be 154% more expensive than the cheapest one that had the poorest states of attributes and was located in the cheapest location attractiveness zone. After the optimisation of the *MSE* for statistical and expert approaches, this ratio was set at the levels of 1.37 and 1.24, respectively.

In order to apply the algorithm (1) to the statistical and expert approaches, weights of attributes had to be calculated. For the expert approach, four real estate appraisers did pairwise comparisons of attributes in the AHP method. The results (AHP matrixes – C_1, C_2, C_3, C_4) are presented in equations (15) – (18) (the order of rows/columns is the same as the order of attributes: area, access to utilities, accessibility of public transport, quality of surroundings and attractiveness of the shape of a plot):

$$C_1 = \begin{bmatrix} 1 & \frac{1}{3} & 1 & \frac{1}{3} & 1 \\ 3 & 1 & 3 & 2 & 3 \\ 1 & \frac{1}{3} & 1 & \frac{1}{3} & \frac{1}{2} \\ 3 & \frac{1}{2} & 3 & 1 & 3 \\ 1 & \frac{1}{3} & 2 & \frac{1}{3} & 1 \end{bmatrix}, \tag{15}$$

$$C_2 = \begin{bmatrix} 1 & 2 & 2 & \frac{1}{2} & 1 \\ \frac{1}{2} & 1 & 4 & 3 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & 1 & \frac{1}{2} & \frac{1}{4} \\ 2 & \frac{1}{3} & 2 & 1 & 2 \\ 1 & 2 & 4 & \frac{1}{2} & 1 \end{bmatrix}, \tag{16}$$

$$C_3 = \begin{bmatrix} 1 & \frac{1}{4} & 2 & \frac{1}{4} & 2 \\ 4 & 1 & 4 & \frac{1}{2} & 4 \\ \frac{1}{2} & \frac{1}{4} & 1 & \frac{1}{3} & 2 \\ 4 & 2 & 3 & 1 & 4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}, \tag{17}$$

$$C_4 = \begin{bmatrix} 1 & 1 & 1 & \frac{1}{3} & 2 \\ 1 & 1 & 1 & \frac{1}{3} & 2 \\ 1 & 1 & 1 & \frac{1}{3} & 2 \\ 3 & 3 & 3 & 1 & 3 \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{3} & 1 \end{bmatrix}. \tag{18}$$

The weights of attributes obtained on the basis of matrixes C_1, C_2, C_3, C_4 are presented in Table 1.

TABLE 1. WEIGHTS OF ATTRIBUTES FOR THE AHP METHOD FOR ALL APPRAISERS

Appraiser	Attributes				
	area	access to utilities	accessibility of public transport	quality of surroundings	attractiveness of the shape of a plot
1	10.91%	37.73%	9.73%	28.73%	12.91%
2	20.27%	24.75%	7.30%	24.33%	23.35%
3	12.36%	31.39%	10.17%	39.11%	6.97%
4	16.09%	16.09%	16.09%	42.26%	9.47%

S o u r c e: author's calculation.

As presented in Table 1, the weights obtained from the four experts were diverse. The highest weights were obtained for the access to utilities and the quality of surroundings, whereas the lowest for the accessibility of public transport and the attractiveness of the shape of a plot. The AHP matrix for the second expert was not consistent (matrixes obtained for the three other appraisers were), despite the fact that this person's assessments were repeated twice (the *consistency ratio* for this expert equalled 1.17). However, the author decided to leave it in the study, because some researches demonstrated that vectors of weights obtained for participants with the consistency ratio higher than 0.1 were not significantly different from these obtained for the consistency ratio equal to or lower than 0.1 (Apostolou and Hassell, 1993). Subsequently, mean weights for each attribute were used in the expert approach. The weights of attributes for the statistical approach (based on the partial Kendall τ_B coefficients) and for the expert approach are presented in Table 2.

TABLE 2. WEIGHTS OF ATTRIBUTES FOR STATISTICAL AND EXPERT APPROACHES

Approach	Attributes				
	area	access to utilities	accessibility of public transport	quality of surroundings	attractiveness of the shape of a plot
Statistical	7.57%	—	51.42%	31.78%	9.24%
Expert	14.91%	27.49%	10.82%	33.61%	13.17%

S o u r c e: author's calculation.

Weights obtained by the statistical and expert approaches are quite different from one another. Besides the fact that it was impossible to estimate the weight of the 'access to utilities' attribute for the statistical approach (in the expert approach, the weight of this attribute was the second largest, at 27.5%), the most significant difference can be observed in the case of the 'accessibility of public transport' attribute. The experts associated the least weight to this attribute (10.8%), whereas from the point of view of the strength of the relationship between this attribute and the unit value of the representative properties, calculated by the statistical approach, its weight (over 51.0%) surpassed the sum of the weights of all the other attributes. The 'quality of surroundings' attribute had a similar weight in both approaches. The smallest weights in the statistical approach were obtained for the 'area' and the 'attractiveness of the shape of a plot' attributes. They also assumed relatively small weights in the expert approach. These weights were transformed into the impact of attributes on the unit value of a property using equations (8) and (9).

In the econometric approach, it was impossible to estimate the classical regression model (5) because of the collinearity of attributes. In order to overcome this difficulty, the ridge regression (Tikhonov regularisation) was applied (Calvetti et al., 2000). The regularisation parameter was set at the level of 0.0001. Table 3 presents the vectors of the impact of attributes on the unit value of properties $(1 + a_{kp})$ for the econometric, statistical and expert approaches.

TABLE 3. VECTORS OF IMPACT OF ATTRIBUTES ON UNIT VALUE OF PROPERTIES $(\mathbf{1} + \boldsymbol{A_{KP}})$

Attributes	States of attributes	Approach				
		Ec	St	Ex	St_{MSE}	Ex_{MSE}
Area	1	1.000	1.000	1.000	1.000	1.000
	2	0.987	1.036	1.072	1.012	1.016
	3	1.003	1.073	1.149	1.024	1.033
Access to utilities	1	—	—	1.000	—	1.000
	2	—	—	1.137	—	1.030
	3	—	—	1.292	—	1.061
Accessibility of public transport	1	1.000	1.000	1.000	1.000	1.000
	2	0.969	1.270	1.052	1.085	1.012
	3	1.032	1.614	1.106	1.176	1.024
Quality of surroundings	1	1.000	1.000	1.000	1.000	1.000
	2	1.014	1.104	1.110	1.034	1.025
	3	1.039	1.218	1.232	1.069	1.050
	4	1.097	1.344	1.367	1.106	1.075
Attractiveness of the shape of a plot	1	1.000	1.000	1.000	1.000	1.000
	2	1.083	1.044	1.063	1.015	1.014
	3	1.058	1.090	1.131	1.030	1.029

S o u r c e: author"s calculation.

Analysing the vectors of impact of the states of attributes on the unit value of a property, we can see that the highest values of $1 + a_{kp}$ in the statistical and expert approaches (both with and without the minimisation of the MSE) were

obtained for the attributes with the highest weight and the lowest otherwise (compare Table 2). Because the $\frac{v_{\max}}{v_{\mathrm{b}}}$ ratio for non-optimised approaches was higher than the ratio for the optimised ones, the values of $1 + a_{kp}$ in the former approaches were higher. A higher $\frac{v_{\max}}{v_{\mathrm{b}}}$ ratio will certainly cause more significant dispersion of results. When we compare the results of the statistical and expert approach with the econometric one, it turns out that the latter has most in common with the statistical approach with the minimisation of the MSE, but anyway the differences between the two approaches are relatively significant. The first observation is that in the econometric approach, higher states of attributes do not always have to have a stronger impact on the unit value of a property. In fact, this is only true in the case of one attribute, namely the quality of surroundings. This results from the fact that the values of properties estimated by experts have not always been higher when the states of attributes were higher. Moreover, the econometric approach considers all the states of each attribute separately (as dummy variables do), while in the statistical approach, the correlation strength and direction are reflected for the whole range of the attribute.

After the application of the above-presented results to the whole set of 318 properties, the measures of the accuracy of valuations were obtained (Table 4).

TABLE 4. MEASURES OF ACCURACY OF VALUATIONS (THE MOST ACCURATE VALUES ARE BOLDED)

Measures of accuracy	Approach				
	Ec	St	Ex	St_{MSE}	Ex_{MSE}
MPE	**0.77%**	–0.87%	–2.18%	0.92%	0.79%
$MAPE$	4.74%	7.96%	7.82%	**4.51%**	5.21%
$RMSE$	36.39	59.55	56.53	**35.94**	40.19
Minimum appraised value	**520.19**	427.25	419.70	522.69	543.98
Maximum appraised value	644.97	729.09	**723.90**	631.46	619.53
Maximum underestimation	–14.54%	–23.10%	–21.32%	**–9.71%**	–14.24%
Maximum overestimation	16.72%	25.27%	30.19%	**14.45%**	15.53%
Share of properties with errors between ±5%	57.23%	38.37%	36.16%	**58.49%**	53.15%
Share of properties with errors between ±10%	88.37%	70.76%	66.35%	**93.08%**	84.59%
Share of properties with errors between ±15%	99.37%	86.16%	87.42%	**100.00%**	99.69%

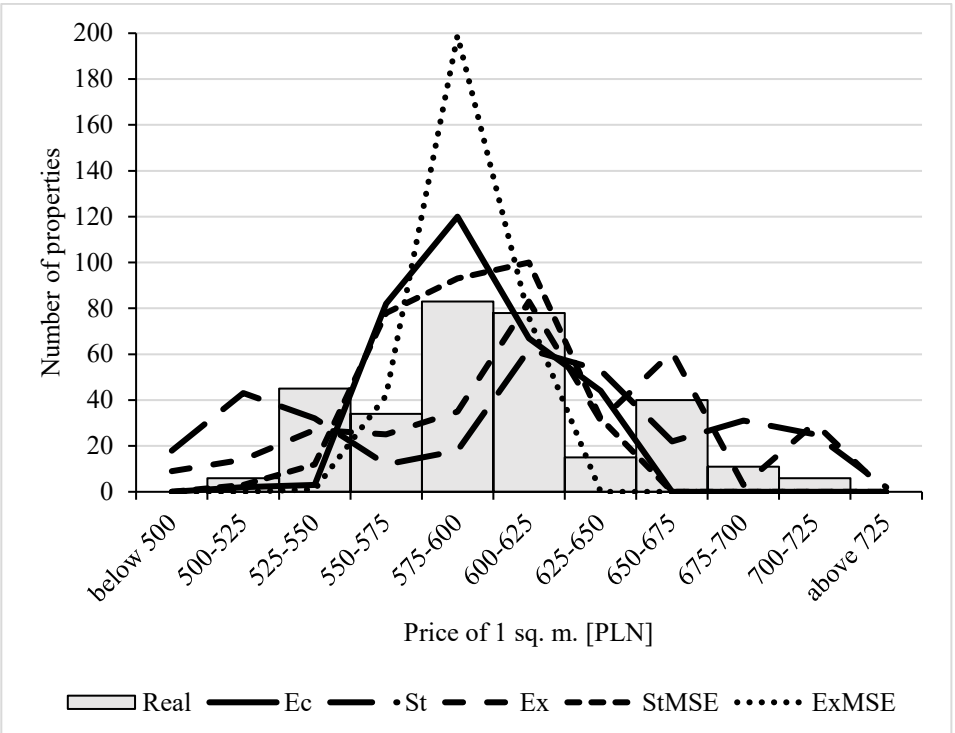
S o u r c e: author's calculation.

It can be seen that as regards the accuracy of valuations, we can divide applied methods into two groups. The first, with relatively high accuracy, consists of the econometric approach and statistical and expert approaches with the minimisation of the MSE. The second group, presenting lower accuracy, comprises

statistical and expert approaches without the minimisation of the MSE. In the first group, the valuations tend to be slightly underestimated (the smallest mean percentage error – 0.767%, was obtained for the econometric approach). In the second group, the situation is opposite – valuations are overestimated, with the mean overestimation equal to 0.87% for the statistical approach and 2.18% for the expert approach. The smallest mean absolute percentage and standard errors were obtained for the statistical approach with the minimisation of the MSE. If we compare the extreme (minimum and maximum) valuations with the real values (502.11 and 701.43 PLN), the minimum valuation closest to the real valuation was obtained for the econometric approach, while the maximum valuation closest to the real one – for the expert approach. All the remaining measures of accuracy indicate that the most accurate mass real estate appraisal approach is the statistical approach with the minimisation of the MSE. In this approach, the maximum underestimation of a unit value of the appraised properties was below 10%, and the highest overestimation – less than 15%. Almost 58.5% of all the valued properties had the valuation percentage error equal to or lower than ±5%. Valuation errors lower than ±10% were obtained for over 93% of all the properties, and the highest valuation error equalled ±15%. As expected, higher range of valuations was obtained for the statistical and expert approaches without the minimisation of the MSE than for the variants with the minimisation of the MSE (because of the higher $\frac{v_{\max}}{v_{\mathrm{b}}}$ ratio in the former). It is worth noting that even in the group of approaches yielding the least accurate results (statistical and expert approaches without the minimisation of the MSE), the results are still relatively accurate. The maximum relative errors oscillate around 30%, over 86-87% of all properties had relative valuation errors smaller than ±15%, and the mean relative error did not exceed 8%.

The results presented in Table 4 demonstrate that in the analysed case, the most effective approach among the methods for mass real estate appraisal was the statistical approach with the minimisation of the MSE, closely followed by the econometric approach. The expert approach with the minimisation of the MSE was the third most effective method, whereas the statistical and expert approaches without the minimisation of the MSE were the least effective ones. However, assessing the quality of methods on the basis of only a few parameters might be misleading. If we look at the distribution of unit values of properties obtained by means of the analysed approaches and the real values, we can observe that the expert approach with the minimisation of the MSE caused strong narrowing of the value range (Figure 2).

Figure 2. Distributions of the real and the estimated unit values of properties



S o u r c e: author's calculation.

The distributions of unit values of properties obtained using the econometric and statistical approaches with the minimisation of the MSE were relatively close to the real ones, but the extreme values tended to be over- or underestimated. The extreme values, on the other hand, turned out to be estimated relatively precisely by the statistical and expert approaches without the minimisation of the MSE. This demonstrates that the minimisation of the MSE is a compromise: its application decreases average valuation errors, but at the same time causes flattening of the results. This particular behaviour of the results was useful for this study, but it will not necessarily be useful in all the cases. Therefore, it should always be decided whether the advantages of the optimisation of valuation errors overweight the disadvantages (narrowing of the results, underestimation of high values and overestimation of low values of properties).

5. CONCLUSIONS

The aim of the article is to perform real estate mass appraisal using decision-making methods, and then to compare the obtained results with the results of the classical (econometric, statistical and expert) approaches. The outcome of

the study demonstrates that adopting decision-making methods (in this case, the quadratic programming) greatly improves the accuracy of estimations calculated by means of the Szczecin Algorithm of Real Estate Mass Appraisal (SAREMA). The statistical approach, based on the partial Kendall τ_B correlation coefficients with the minimisation of the MSE, turned out to be the most accurate one, followed closely by the econometric approach. The expert approach with the minimisation of the MSE, although yielding relatively precise results, was noticeably less accurate than the two above-mentioned approaches, which just confirmed the earlier expectations. It is because both the statistical approach with the minimisation of the MSE and the econometric approach are based on the real relationships between the attributes of a property and its value per 1 m^2. The expert approach, based on real estate appraisers' experience, on the other hand does not always reflect the real relationships in the real estate market. Therefore it could be recommended as a complementary approach, adopted when the statistical or econometric approaches could not be used (e.g. because of small or incomplete data sets).

Although the statistical approach with the minimisation of the MSE is the most effective method in the presented research, it cannot be universally acclaimed as the best one. Relatively much depends on the quality of the used database, which, in turn, is determined by the quality of work of experts (real estate appraisers). If they define the attractiveness zones correctly (properties belonging to each attractiveness zone are alike, i.e. for given states of attributes their values are similar), if the real estate attributes are correctly identified and the representative properties are properly appraised (their value is higher for better states of attributes and lower for poorer states of attributes), then the quality of database could be regarded as good. In such a case, the base value would reflect the value of a property in the cheapest location attractiveness zone with the poorest values of attributes, and the maximum value would represent the value of a property in the most expensive location attractiveness zone and with the best values of attributes. A database prepared with such a degree of diligence would not even need the optimisation of valuation errors, because the original base and maximum values would reflect the variability of real estate values accurately.

It should also be stressed that even though the above findings apply to the analysed 318 properties, they cannot be automatically employed to every situation. Before the application of any of the above-mentioned methods to other pieces of research, it should first be checked which approach would generate the best results for that particular case. It is generally hard to find a method which would be universally applicable to the real estate market, because every territorial unit, i.e. every city, town or village has its own specificity, to which the applied methods should be adapted.

Although the SAREMA has been used for 20 years, this study introduces new approaches to the calibration of the influence of attributes for this algorithm (sta-

tistical and expert approaches with the minimisation of valuation errors), presents their advantages and disadvantages, and facilitates the selection of the most relevant approach for a particular study. Further research into this subject should cover other methods of calibrating the influence of attributes (where transitions between the states would not necessarily be linear) and other approaches to the applications of the SAREMA.

REFERENCES

Apostolou B., Hassell J. M., (1993), An empirical examination of the sensitivity of the analytic hierarchy process to departures from recommended consistency ratios, *Mathemmatical. Computer Modelling*, 17(4/5), 163–170. DOI: 10.1016/0895-7177(93)90184-Z.

Ball J., Srnivasan V. C., (1994), Using the Analytic Hierarchy Process in House Selection, *The Journal of Real Estate Finance and Economics*, 9, 69–85. DOI: 10.1007/BF01153589.

Brunelli M., (2015), Introduction to the Analytic Hierarchy Process, Springer, Cham Heidelberg, New York, London. DOI: 10.1007/978-3-319-12502-2.

Calvetti D., Morigi S., Reichel L., Sgallari F., (2000), Tikhonov regularization and the L-curve for large discrete ill-posed problems, *Journal of Computational and Applied Mathematics*, 123(1–2), 423–446. DOI: 10.1016/S0377-0427(00)00414-3.

Ćetković J., Lakić S., Lazarevska M., Žarković M., Vujošević S., Cvijović J., Gogić M., (2018), Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application, *Complexity*, (special issue, article ID 1472957), 1–10. DOI: 10.1155/2018/1472957.

Dmytrów K., Gnat S., Kokot S. (2018), Próba uwzględnienia kształtu nieruchomości gruntowych jako atrybutu w procesie zalgorytmizowanej wyceny, *Problemy Rynku Nieruchomości*, 2(50), 4–10.

Doszyń M., (2017), Statistical Determination of Impact of Property Attributes for Weak Measurement Scales, *Real Estate Management and Valuation*, 25(4), 75–84. DOI: 10.1515/remav-2017–0031.

Doszyń M., (2018), Ekonometryczna wersja szczecińskiego algorytmu masowej wyceny nieruchomości, *Problemy Rynku Nieruchomości*, 50(2), 11–17.

Foryś I., Gaca R., (2016), Application of the Likert and Osgood Scales to Quantify the Qualitative Features of Real Estate Properties, *Folia Oeconomica Stetinensia*, 16(2). 7–16. DOI: 10.1515/foli-2016-0021.

Guitouni A., Martel J. M., (1998), Tentative guidelines to help choosing an appropriate MCDA method, *European Journal of Operational Research*, 109(2), 501–521. DOI: 10.1016/S0377-2217(98)00073-3.

Hozer J., Kokot S., Foryś I., Zwolankowska M., Kuźmiński W., (1999), *Ekonometryczny algorytm masowej wyceny nieruchomości gruntowych*, Uniwersytet Szczeciński, Szczecin.

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

Hwang C. L., Yoon K., (1981), *Multiple Attribute Decision-Making: Methods and Applications*, Springer-Verlag, New York.

Kauko T., D'Amato M., (eds.), (2008), *Mass Appraisal Methods: An International Perspective for Property Valuers*, Blackwell Publishing, Oxford.

Kolenda M., (2006), *Taksonomia numeryczna. Klasyfikacja, porządkowanie i analiza obiektów wielocechowych*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Kozioł-Kaczorek D., (2012), Hierarchizacja cech nieruchomości z zastosowaniem analitycznego procesu hierarchicznego, *Studia i Materiały Towarzystwa Naukowego Nieruchomości*, 20(1), 165––174.

Nermend K., (2017), *Metody analizy wielokryterialnej i wielowymiarowej we wspomaganiu decyzji*, Wydawnictwo Naukowe PWN, Warszawa.

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. DOI: 10.1016/j.beth. 2010.08.006.

Saaty T. L., (1980), *The Analytic Hierarchy Process*, McGraw-Hill, New York.

Saaty T. L., (1990), How to make a decision: The analytic hierarchy process, *European Journal of Operational Research*, 48(1), 9–26. DOI: 10.1016/0377-2217(90)90057-I.

Saaty T. L., Ergu D., (2015), When is a Decision-Making Method Trustworthy? Criteria for Evaluating Multi-Criteria Decision-Making Methods, *International Journal of Information Technology & Decision Making*, 14(6), 1171–1187. DOI: 10.1142/S021962201550025X.

Triantaphyllou E., (2000), *Multi-criteria decision-making methods: A comparative study*, Kluwer Academic Publishers, London. DOI: 10.1007/978-1-4757-3157-6.

Trzaskalik T., (ed.), (2014), *Wielokryterialne wspomagania decyzji. Metody i zastosowania*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Yalpir S., (2014), Forecasting residential real estate values with AHP method and integrated GIS, in: *Conference proceedings of People, Buildings and Environment, an international scientific conference,* Kroměříž, Czech Republic, 694–706.

Źróbek S., Bełej M., (2000), *Podejście porównawcze w szacowaniu nieruchomości*, Educaterra, Olsztyn.

DOI: 10.5604/01.3001.0014.0954

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 4 – 2019

Czesław DOMAŃSKI[1]

The birth of official statistics in Poland

Abstract. *The birth of the Polish official statistics, commonly associated with the establishment of Statistics Poland on 13 July 1918, was possible thanks to the ideas and grassroots work of various scholars, including statisticians, who thus demonstrated their faithfulness to the resolutions of the Four-Year Sejm (Parliament) and the Constitution of 3 May 1791. Statisticians from Kraków, who established Poland's first statistical society (the Polish Statistical Society) and authored a study entitled* Statistics of Poland*, Kraków (1915), edited by Adam Krzyżanowski and Kazimierz Kumaniecki, deserve a special recognition in this paper. In the two consecutive years, 1916 and 1917, Kraków saw the publication of two important works, namely* Geographical and Statistical Atlas of Poland*, Kraków (1916), edited by Eugeniusz Romer and Ignacy Weinfeld, and the* Yearbook of Poland. Statistical Tables*, Kraków (1917). But scientists from other parts of Poland contributed to the process of laying the foundations for the Polish official statistics as well. Statisticians from Warsaw compiled three volumes of* Statistical Yearbooks of the Kingdom of Poland *edited by Władysław Grabski, and the work entitled* Statistical Yearbook of the Kingdom of Poland and other Polish lands. The Year 1915 *under the editorship of Edward Strasburger was published in 1916 in Sankt Petersburg. Around the same period, Polish scientists associated with the University of Lviv expanded the scope of their scientific activity. After regaining independence, the process of establishing the system of official statistics in Poland, at that time carried out already within the framework of Statistic Poland, continued developing dynamically thanks to the scientific input of several dozen Polish statisticians.*

Keywords: official statistics, statistical societies, precursors of public statistics

Narodziny statystyki publicznej w Polsce

Streszczenie. *Narodziny polskiej statystyki publicznej, utożsamiane z utworzeniem Głównego Urzędu Statystycznego 13 lipca 1918 r., mogły nastąpić dzięki ideom pracy u podstaw wielu uczonych, w tym statystyków, którzy byli wierni postanowieniom Sejmu Czteroletniego i Konstytucji 3 Maja 1791 r. Na szczególne wyróżnienie zasługują tutaj statystycy krakowscy, którzy powołali do życia pierwsze Polskie Towarzystwo Statystyczne, oraz ich dzieło* Statystyka Polski *pod redakcją Adama Krzyżanowskiego i Kazimierza Kumanieckiego (Kraków 1915). W następnych dwóch latach zostały wydane dwie ważne prace naukowe:* Geograficzno-Statystyczny Atlas Polski *pod redakcją Eugeniusza Romera i Ignacego Weinfelda (Kraków 1916) oraz* Rocznik Polski. Tablice statystyczne *(Kraków 1917). Ale naukowcy z innych stron Polski również wnieśli wkład w proces tworzenia*

[1] University of Lodz, Faculty of Economics and Sociology, Department of Statistical Methods, ul. Rewolucji 1905 r. 41, 90-214 Łódź; Poland, e-mail: czeslaw.domanski@uni.lodz.pl, ORCID: https://orcid.org/0000-0001-6144-6231.

podwalin polskiej statystyki publicznej. Statystycy warszawscy opracowali trzy tomy Roczników Statystycznych Królestwa Polskiego *pod redakcją Władysława Grabowskiego, natomiast w 1916 r. w Petersburgu wydany został* Rocznik Statystyczny Królestwa Polskiego z uwzględnieniem innych ziem polskich. Rok 1915 *pod redakcją Edwarda Strasburgera. W tym okresie również rozwinęła się twórczość polskich uczonych związanych z Uniwersytetem Lwowskim. Proces organizacji statystyki publicznej w Polsce po odzyskaniu niepodległości, już w ramach Głównego Urzędu Statystycznego, mógł rozwijać się w dynamicznym tempie dzięki pracy naukowej kilkudziesięciu statystyków polskich.*

Słowa kluczowe: statystyka publiczna, towarzystwa statystyczne, prekursorzy statystyki publicznej

JEL Classification: B29, C1, J16

1. INTRODUCTION

The establishment of Statistics Poland on 13 July 1918 was possible thanks to ideas shared by a group of scholars who bore in mind the resolutions of the Four Year Sejm (the Parliament) and the Constitution of 3 May 1791. Following the message of these documents, soon after the last partition of Poland in 1795, several scientific societies came to existence in the former territory of Poland, whose programmes propounded the idea of conducting research into the local population structure and economy.

The Warsaw Society of the Friends of Sciences[2], which assembled the most distinguished Polish scientists of the post-partition era, was founded in 1800. In 1807 in Kraków, Stanisław Staszic published the work *On Polish Statistics*, which was the response to launching statistical surveys throughout the country. The first statistical data on the Duchy of Warsaw was published in 1809 by Jerzy B. Flatt in the work entitled *The description of the Duchy of Warsaw,* and the first Statistical Office was established within the Ministry of the Interior in the Duchy of Warsaw in 1810 (Kubiczek, 2017).

In the first two decades of the 20th century, two statistical scientific societies were established: the Polish Statistical Society, in 1912, and the Society of Polish Economists and Statisticians, in 1917. Both of them became the pillars of the foundation of Statistics Poland (Domański, 2008). However, the origins of Statistics Poland could also be traced back to the statistical institutions created by the partitioners; as it was them who provided the sources of statistical data, launched studies on the Polish territories and provided experienced statisticians ready to organize statistical services in the independent Polish state. The National Statistical Bureau was established in 1873 in Lviv, a year after the Lviv Municipal Statistical Office started operations, and in 1884, a similar municipal-statistical unit opened in Kraków.

[2] The Warsaw Society of Friends of Science (Polish: *Towarzystwo Przyjaciół Nauk*, *TPN*) was one of the earliest Polish scientific societies, active in Warsaw from 1800 to 1832.

In the Kingdom of Poland, the Government Commission of Internal and Clerical Affairs, within which the Statistical Department dealing with collecting numerical data operated, was established in 1815. Another significant statistical organisation from the territory of the Russian partition was the Warsaw Statistical Committee, which functioned in 1889–1914 and published, among other works, a 40-volume series of *Works of the Warsaw Statistical Committee* (Łazowska, 2015).

In 1827 in Poznań, Stanisław Plater published the first *Statistical Atlas of Poland,* which defined the territorial scope of the Polish culture and nationality and presented the distribution of industry in the Polish territories. The Poznań Society of Friends of Sciences, composed of Polish statisticians who established the Faculty of Economic and Statistical Sciences in the 1870s, came into being in 1857. The Faculty later merged with the Faculty of Economic Sciences. Stanisław Plater continued his scientific activity until the outbreak of World War I. The Statistical Office in Poznań, which generated a much broader scope of regional statistics than its predecessor, Berlin-based Royal Prussian Statistical Office from 1805, was established in 1905 (Łazowska, 2018b).

This paper focuses on the processes leading to the launch of Polish official statistics in two parts of the former Commonwealth of Both Nations (Rzeczpospolita Obojga Narodów): the Polish Kingdom[3] and Galicia. Polish statisticians from those regions were the first and the most active in this process. In other parts of the Polish Commonwealth (Prussian partition and Russian annexed territories – Lithuania, Belarus, Ukraine), Polish official statistics practically did not exist.

2. FIRST SCIENTIFIC SOCIETIES IN POLAND

The beginning of Poland's servitude marks the start of the period when numerous Polish scientific societies were established. The two literary and scientific centres of the post-partition era were the town of Puławy[4], which belonged to the Czartoryski family[5], and Vilnius. However, it was Warsaw where the intellectual life of the time was flourishing. The homes of Joachim Chreptowicz, Tadeusz Czacki, Ignacy Krasicki, and finally Anna nee Sapieha and Stanisław Sołtyk, were the places where the intellectual elite of the former capital met. The discussions and debates held there among the Warsaw intellectuals gave rise to

[3] The Kingdom of Poland, Congress Poland or Russian Poland was a polity created in 1815 by the Congress of Vienna as a sovereign Polish state.

[4] A city in eastern Poland, in Lublin Province of northern Lesser Poland, located at the confluence of the Vistula and Kurówka rivers. In 1731, Maria Zofia Sieniawska (the daughter of Elżbieta and Adam Sieniawski), married August Aleksander Czartoryski. As a result, Puławy remained in the hands of the Czartoryski family for the next 100 years.

[5] Czartoryski is a Polish princely family of Lithuanian-Ruthenian origin, also known as the Familia.

the idea of creating a more permanent scientific association. The man who is believed to be the main originator and propagator of the idea was Stanisław Sołtyk.

Sołtyk's idea was implemented in 1800 in the form of the Warsaw Society of the Friends of Sciences, which attracted the most prominent scientists of the post-partition era. The main aims of the Society were to support the development of sciences and to promote studying in Polish, except for the 'matters related to home religion and the present government', i.e. the most sensitive political issues of the time. The Society became a platform for the presentation of the scientific work of economists, at the same time serving as a source of inspiration for them to undertake new challenging tasks. Numerous discussions on socio-economic problems were held there. In 1804, Stanisław Staszic conducted the analysis of Aleksander Potocki's work on agriculture, and in December the following year he presented the first part of his *Agriculture of the Carpathian.* Soon after that Staszic became the informal leader of the Society, and from 1808 until his death in 1826 he served as its president, throughout that time generously supporting it financially. The society could boast many active members, including Dominik Krysiński, who made his debut in 1807 giving a talk on tanning, or Wawrzyniec Surowiecki, who in 1811 gave a lecture on the decline of Polish cities.

In 1818, Surowiecki presented his doctoral dissertation on trade guilds before the Society. The subsequent discussion on guilds, which was joined, among others, by Dominik Krysiński, became a valuable contribution to the wide-scale debate on the extra-economic factors of the social development, held at that time by Polish physiocrats.

Fryderyk Skarbek, who was the author of the contemporarily published *Elementary Principles of the National Economy*, joined the Society in 1820. In the following years, he proved to be one of the most active members, contributing his expertise in areas such as the Galician statistics, charitable institutions, providence and saving funds for farmers.

After the death of Staszic in 1826, Julian Ursyn Niemcewicz took the lead of the Society. Studies in the field of economics were still the members of the Society's main focus. Dominik Krysiński's dissertation on the national economy and its analysis carried out by Skarbek became the highlight of the year 1826 (Łukasiewicz, 1995).

The Society remained indifferent to the social unrest which preceded the outbreak of the November Insurrection of 1830. The defeat of the Insurrection had several negative political, economic and military consequences. One of them was the liquidation of the Warsaw Society – a severe loss to Polish science. The last public meeting of the Society was held on 3 May 1831, when Tsar Mikolai I decided to confiscate all its assets. That event marked the beginning of the period of the so-called Paskiewicz terror, which forced many eminent members of the Society into exile.

The Warsaw Society of Friends of Sciences played an important role in the formation of the modern Polish science. In addition to bringing together several outstanding scientists and serving as a platform for exchanging scientific ideas, the society published yearbooks and motivated scientists to conduct further research. The Society's members collected national documents and important diaries, and preserved them in the institution's library, which became a respected research unit.

In 1818, the Society of Friends of Sciences of the Lublin province was founded in Lublin on the initiative of a landowner Joachim Powidzki. The Lublin Society, which was guided by the principles of the Enlightenment, 'focused on efforts of social activists rather than scientists'.

In the Płock province, intensive research studies were carried out by the Płock Scientific Society, founded in 1820 by Kajetan Morykoni, the President of the Provincial School. Unfortunately, after the November Insurrection, both of the above-mentioned Societies shared the fate of the Warsaw Society (Roszkowski, 2010).

The Society of Support for Social Work (TPPS), an association through which economists, lawyers and economic and political activists carried out joint social activities, was established in Warsaw in 1907. The Society's goal was to develop scientific research, compile technical data, prepare drafts of bills and hold discussions on socio-economic problems.

The founders of the Society included Franciszek Nowodworski, Zygmunt Chrzanowski, Leopold Kronenberg, Juliusz Tarnowski and Władysław Grabski. In 1910, a scientific unit, under the name of the Social Work Office, was set up within the Society (TPPS). Its aim was to provide scientific support to the Polish representatives in the National Duma (Russian parliament). Initially, there were two sections in the Office: the socio-legal unit and the socio-economic one, headed by Władysław Grabski (1874–1938). Several well-known economists participated in the works of this section, including Stanisław Bukowiecki, Bohdan Wasiutyński, Stanisław Aleksander Kempner, Henryk Radziszewski and Bolesław Markowski. Their work involved a wide range of subjects, for example the problems of municipal and land self-governments, agrarian issues, local finances and tax and communication issues. In the years 1910–1915, two statistical yearbooks of the Kingdom of Poland for the years 1913 and 1914, edited by Władysław Grabski, were published. The first of them was issued in January 1914 under the title of the *Statistical Yearbook of the Kingdom of Poland. The Year 1913* thanks to a donation from 'Dr Józef Mianowski Provident Fund'. The data collected in that yearbook was classified under the following headings: area and population including emigrations; property and agricultural production; industry, trade and credits; constructions and fires; roads; taxation and urban households; public health; education.


Rocznik Statystyczny
Królestwa Polskiego
Rok 1914

Similar categories were adopted in the second edition of the yearbook for the year 1914 (cf. photograph), also edited by professor Władysław Grabski, and published the following year.

In the third edition of the yearbook, for the year 1915, both the thematic and territorial scopes of the presented data were considerably expanded (by adding numerical data on the Polish territories beyond the Kingdom of Poland). The publication, edited by Edward Strasburger, came out in Warsaw and in Sankt Petersburg in 1916 under the title *Statistical Yearbook of the Kingdom of Poland and other Polish Lands. The Year 1915.*

3. POLISH STATISTICAL SOCIETIES

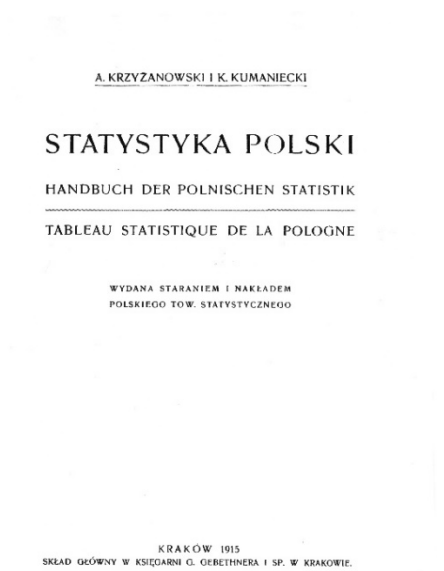
The beginnings of the Polish statistics date back to the end of the 19th century, the period marked by the development of the Polish scientific thought in the field of socio-economic sciences in Galicia. In 1867, the Law and Economics Society, whose task was to educate and develop jurisprudence and socio-economic sciences, was established in Kraków. The Trade and Economics Society, founded to support trade and to collect economic data on the countries-destinations of Poles' emigration, started operations in 1894 in Lviv.

The idea to found a separate association for Polish statisticians originated at the beginning of 1912 in the circles of statisticians and economists from Kraków. They planned that this future association would issue statistical publications covering the joint territory of the three partitions. The first president of the Polish Statistical Association was Prof. Juliusz Leo, an expert in administrative law, who later became the Mayor of Kraków, while Prof. Kazimierz Kumaniecki, a statistician and lawyer from the Jagiellonian University, served as the secretary (Berger, 2008; Kruszka, 2012).

However, publications covering the joint territory of the three partitions could not be issued until 1914; it was mainly due to the lack of uniformity of statistical data collected by the authorities of the partitioners. Statistical data regarding Polish territories were scattered through several publications issued separately in each of the partitioning countries. It was possible to compile the data from these various publications in one yearbook, but only after those data had been made comparable, which was a task for an institution that would be independent of the state units of administrative statistics (Berger, 1995).

According to its statute, the main aim of the Polish Statistical Society was 'statistical research on the economic and social relations between the Polish territo-

ries and the Polish emigration'. The Polish Emigration Society operated in Kraków at that time, but it did not conduct any separate statistical research. Therefore, the Polish Statistical Association treated that task as one of its priorities. In order to achieve these goals, there were plans to establish the Statistical Office for the Polish Territories, whose responsibility would be publishing statistical yearbooks, statistical monographs and handbooks of statistics, as well as collecting all the necessary input, including statistical data obtained through appropriate questionnaires, and, as mentioned in the statute, through 'on the spot' research. A special commission, whose members were to be appointed by the PSS, was to deal with these issues. The ambitious plan to launch a wide-scale publishing activity came to fruition in 1915, with the publication of *Statistics of Poland* by the Jagiellonian University (cf. photograph). The compilation and publication of the yearbook was possible thanks to the endeavours of several institutions from Kraków. In July 1914, the Polish Statistical Society had to apply for financial support necessary to conduct studies and finalise the work on the yearbook to the Academy of Learning in Kraków.


A. KRZYŻANOWSKI I K. KUMANIECKI

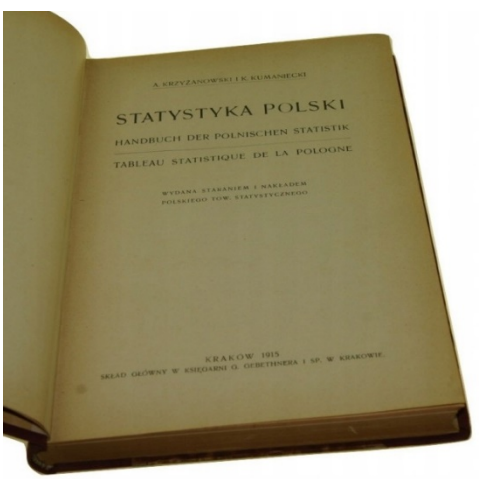
STATYSTYKA POLSKI

HANDBUCH DER POLNISCHEN STATISTIK

TABLEAU STATISTIQUE DE LA POLOGNE

WYDANA STARANIEM I NAKŁADEM
POLSKIEGO TOW. STATYSTYCZNEGO

KRAKÓW 1915
SKŁAD GŁÓWNY W KSIĘGARNI G. GEBETHNERA I SP. W KRAKOWIE.



The authors of *Statistics of Poland* were Professor Adam Krzyżanowski, Ph.D., (1873–1953) and Assistant Professor Kazimierz W. Kumaniecki Ph.D., (1880–1941). The Editorial Committee, selected by the Board of the PSS, consisted of Professor Franciszek Bujak Ph.D., (1875–1953) (the head of the Committee), Professor Stefan Surzycki Ph.D., (1864–1926), and Edward Grabowski Ph.D., (1881–1961), who authored the chapter devoted to the sources of statis-

tical data. Among the important contributors, Marcin Nadobnik Ph.D., (1883––1953) from Lviv and Michał Römer Ph.D., (1880–1945) from Vilnius should be mentioned, as they supported the publication with their own statistical materials. Other contributors included professor Władysław Leopold Jaworski Ph.D., (1865–1930), Professor Stanisław Kutrzeba Ph.D., (1876–1940), Professor Michał Rostworowski Ph.D., (1864–1940) and Franciszek Stefczyk Ph.D., (1861–1924). The preparation of tables where data were compiled was done by the staff of the Municipal Statistical Office in Kraków (Domański, 2008).

In the preface to *Statistics of Poland*, the Editorial Committee defined the aim of the publication in the following way: 'Our work was resolved with a handbook in mind. We neither intended to provide a complete compilation of statistical figures, nor offer a publication which would make the use of statistical sources redundant. Our intention was rather to show the sources and to facilitate the use of extensive statistical material by presenting the most important calculations in a concise way'.

Statistics of Poland also attempted to compile numerical data concerning the territory of the pre-partitioned Poland; however, the thematic, chronological and territorial scope of the publication was not uniform, as it largely depended on the availability of reliable statistical sources. This fact by no means lessens the influence of the publication on the development of the Polish statistical thought. *Statistics of Poland* was Poland's first statistical yearbook in history which encompassed statistical data from the three partitions. The data presented there fell under categories such as: population, economic and social relations, education and learning and elections to legislative bodies. The publication was illustrated with a series of supplementary tables. Both the hand-written and printed materials have been used by many researchers who have studied the historical and statistical issues related to the territory of Poland, for instance by Professor Eugeniusz Romer, Ph.D., who used them in his work entitled *Geographic and Statistical Atlas of Poland,* published in 1916.

Thanks to the publication of *Statistics of Poland* in 1915, the Polish Statistical Society fully achieved its statutory aims. It successfully created the first yearbook of the Polish territories, thus assuming the responsibilities of the non-existent body of official statistics. *Statistics of Poland* consists of four chapters.

Chapter one, under the 'Population' heading, presents data related to the population numbers and structures according to sex and age, natural and migration movements of population, faith, nationality and language. Chapter two, most substantial in terms of the content and volume, entitled 'Economic and Social Relations', presents tables compiling data on occupation, agriculture and agricultural property, sowing and crops, animal rearing agricultural industry, real estate, constructions, insurance against fire, land indebtedness, land parcelling, industry and trade, mining and metallurgy, strikes, working class associations, social insurance, health service and hospitals, banking institutions and co-operative societies and savings and communications. Chapter three, devoted to education

and schooling, features statistical data on types of schools and their location, number of students and the schooling of children with special educational needs and illiteracy. In chapter four, entitled 'Elections to legislative bodies', the authors present data concerning elections to the Russian Duma, Austrian and German parliaments and the Sejm (parliament) of Galicia.

The publication of *Statistics of Poland* should be viewed nowadays as an original scientific achievement, whose ambitious and innovative goal was to integrate data from the three partitions, at the same time taking into account their various organizational structures, assessment methods, and, consequently, different ways of the compilation and interpretation of the results.

Scientific circles of statisticians and economists from Kraków established the Polish Statistical Society. At that time, Polish scientists tended to associate around several scientific centres throughout the post-partitioned Poland, and channel their scientific activity through these organisations. In Lviv, professor Eugeniusz Romer (1871–1954) wrote *Geographic and Statistical Atlas of Poland* (1916), which played an important role when the boundaries of Poland were delineated after World War I. The materials for that publication came from the library of the Central Statistical Commission, the Chamber of Trade and Industry, the Court and University Library in Vienna, the Jagiellonian University, the Kraków Municipal Office Library, and from private collections of Professor F. Bujak, Rev. J. Fijałek and S. Surzycki from Kraków.

Expert knowledge on the specialised subjects listed below was provided by:

- Professor K. Nitsch (Kraków) – language relations;
- J. Nowak, Ph.D. (Lviv–Vienna) – geology;
- S. Weigner, Ph.D. (Cracow–Vienna) – mining;
- B. Gubrynowicz, Ph.D. (Lviv–Vienna) – Polish press;
- J. Rutkowski, Ph.D. (Warsaw–Vienna) – history;
- W. Szafer, Ph.D. (Lviv–Kielce) – flora.

In his preface to the publication, professor Eugeniusz Romer wrote: 'Let this illustration of speech of numbers on Poland teach our own people and bring deliberateness and goodwill in those who hold in their hands the future of the Polish cause. After all, it is numbers which show how to rule the world' (Romer, 1916, p. 5).

4. THE SOCIETY OF ECONOMISTS AND STATISTICIANS OF POLAND

As has been mentioned before, the economic section was set up within the Social Work Office in 1910, and Zygmunt Chrzanowski become its head. In 1917, when Włodzimierz Wakar assumed the lead of the Office, the section started to prepare statistical materials designed to meet the needs of the future Polish state. Several important documents were prepared and published; for

example, a draft of the Constitution, a draft of the electoral law for the Sejm and the *Polish territorial programme*, which, along with an extensive dissertation by Wakar, became the basis for Poland's stance regarding the border issue at the Versailles Conference. The Office published *Social Work* – its flagship journal, where, among others, Tadeusz Szturm de Sztrem presented materials on the movement of prices and payments in the occupied Warsaw. It was in the circles of the Warsaw scientists where the idea of establishing societies of Polish statisticians and economists emerged.

On 29 March 1917, the Organizing Commission, consisting of Jan Dmochowski, Franciszek Doleżal, Stefan Dziewulski, Kazimierz Kasperski, Stanisław Aleksander Kempner and Włodzimierz Wakar, was set up in Warsaw. The aim of the Commission was to hold a conference which would lay the foundations for the establishment of the Society of Polish Economists and Statisticians in Warsaw, and help obtaining the legalisation of its draft statute from the German occupational authorities. The first meeting of the Society took place in Warsaw on 3 December 1917, and was attended by 42 most distinguished economists and practitioners from Warsaw.

That first meeting named Antoni Kostanecki the chairman of the Society and Stefan Dziewulski and Ludwik Krzywicki deputy chairmen. Jan Dmochowski and Kazimierz Kasperski were chosen for secretaries, and Władysław Zawadzki became the treasurer. The remaining members of the Board were: Zdzisław Ludkiewicz, Jerzy Michalski and Edward Strasburger. The statutory aim of the Society was 'raising the standard and developing the Polish socio-economic knowledge both in terms of theory and practice' by means of publications, talks, discussions, conventions, running a library, organizing research in special areas, announcing competitions and maintaining relations with other similar societies.

At the first meeting of the Board of the Society on 10 December 1917, it was decided that all economic entities from the territory of Poland would have to become registered, that the statistics of all the lectures in the field of socio-economic sciences delivered at Polish universities would be kept, and a syllabus of lectures for vocational schools would be written and sent for approval to the Ministry of Education of the Regency Council[6]. On 30 December 1917, the following meeting of the Society was held, which was devoted to the protection of the Polish economic interests during the peace talks in Brześć Litewski.

The meeting was attended not only by members of the Board of the SPES, but also by the representatives of the government of the Regency Council and the representatives of private business circles. All the conclusions were recorded and sent to the government of the Regency Council along with programme postulates.

[6] The Regency Council of the Kingdom of Poland (Polish: *Rada Regencyjna,* or *Rada Regencyjna Królestwa Polskiego*) was a semi-independent and temporarily appointed highest authority (head of state) in the Partitioned Poland during World War I.

On 14 October 1918, predicting the imminent withdrawal of the occupying forces, the Society devised and announced a plan of the most urgent tasks, the fulfilment of which would enable efficient organisation of the social and economic life in the reunited Poland. The plan involved:

- assessing the economic strengths of particular provinces;
- describing changes caused by the war and occupation;
- adopting a stance concerning the role of foreign capital in Poland;
- assessing the national wealth and income as a basis for the budget;
- establishing a system of taxation;
- defining the forthcoming tasks for the social legislation.

The programme was ambitious and well thought out. This demonstrates that during its first year of operations, the Society of Polish Economists and Statisticians in Warsaw engaged in a number of intensive activities which, due to political reasons, could not be fully overt.

5. BEGINNINGS OF STATISTICS POLAND

On 12 September 1917, the Regency Council was established on the basis of a warrant by two general governors: the German governor Hans von Beseler and the Austrian governor Stanisław Szeptycki. In theory, the newly-established Regency Council was to become the highest authority in the Kingdom of Poland, before handing the power over to a regent or a king.

However, in practice the prerogatives of the Regency Council were much more modest – they were limited mainly to three areas: judiciary, education, and, partly, administration.

The members of the Council appointed by the two emperors were sworn on 15 October 1917. They were:

- Archbishop of Warsaw Aleksander Kakowski;
- Mayor of Warsaw Duke Zdzisław Lubomirski;
- honorary chairman of Real Policy Party Count Józef Ostrowski.

The Regency Council chose the Royal Castle in Warsaw for its headquarters.

In August 1915, Russian authorities left Warsaw hastily, leaving behind statistical materials, archives and contents of the library of the former Warsaw Statistical Commission. In the same month, the entire collection got protection from the Citizen Committee of the City of Warsaw. On 15 September 1915, the care and protection over the collection was entrusted to Professor Ludwik Krzywicki (1859–1941), appointed its custodian (Berger, 2018).

Throughout 1917, the Department of Statistics at the Ministry of Internal Affairs was preparing drafts of laws regulating the organization and activities of the future office of statistics.

Meanwhile, Professor Ludwik Krzywicki was re-appointed as the head of the Department of Statistics at the Ministry of Internal Affairs.

In February the following year, the Department of Statistics was joined by Professor Jan Rutkowski (1886–1949), an economist, historian and later Professor at the Department of Economic History at Poznań University.

As early as February 1918, Professor Ludwik Krzywicki and Professor Jan Rutkowski compiled a draft of the *Project of the Central Statistical Office Organization.*

In March and May 1918, the Project was thoroughly discussed in a round of intra-ministerial conferences. On 11 June 1918, the document, accompanied by an extensive substantiation, was referred to the Council of Ministers, which introduced some amendments to it and passed it on to the Regency Council for approval.

The Project was approved by the Regency Council, which issued a *Decree on the Foundation and Organization of Statistics Poland* on 13 July 1918. The decree provided a legal and organizational framework for the activities of Statistics Poland in the Kingdom of Poland. Moreover, its provisions influenced the shaping of the concept of statistical research in future, i.e. after regaining independence.

The decree of the Regency Council of 8 November 1918 appointed Józef Buzek (1873–1936) the director of Statistics Poland, who remained in this post until 1929 (Łazowska, 2018a).

Professor Józef Buzek, Ph.D., had previously held several important posts connected to statistical research and the organisation of statistical services. He was, among other functions, a member of the Central Statistical Commission in Vienna and a member of the Statistical Council for the Statistical Office of the Austrian Minister of Industry and Trade. In 1907, he was elected a member of the Vienna Parliament, and in 1911, he was re-elected an MP of the Austrian State Council from one of the Lviv constituencies. In 1919, he was transferred from the Austrian Parliament to the Constituent Sejm by the power of the decree issued on 18 November 1918 by the Head of State, Marshal Józef Piłsudski. In the early 1919, he also served as a member of the Polish delegation to the peace conference in Paris.

At first, Statistics Poland was briefly headed by Ludwik Krzywicki (until 8 November 1918), during prolonged job negotiations with Józef Buzek.

6. FINAL REMARKS

When analysing the most important achievements of the Polish scientific societies during Poland's partitions, it can be observed that these societies were very active, despite numerous obstacles created by the partitioning powers, mainly Prussia and Russia. Statisticians representing intellectual circles of Kraków deserve to be distinguished here, as they were the first to establish a Polish scientific society of statistics. They moreover published a fundamental work, *Statistics of Poland*, Kraków, 1915. During World War I, the statisticians from Warsaw compiled two volumes of the *Yearbook of the Kingdom of Poland* (1916, 1917). At the same time, Polish scientists affiliated to the University of Lviv became increasingly active. *The Atlas of Polish Geography and Statistics*, edited by Professor Eugeniusz Romer and Professor Ignacy Weinfeld, was printed in Kraków in 1916. In the following year, Professor Romer compiled the *Yearbook of Poland. Statistical Tables*, Kraków, 1917.

Polish scholars of that time, by compiling statistical data from the territories of three partitions, demonstrated deep understanding of the importance of statistics for the needs of the reborn Poland. A swift and successful launch of public statistical services in Poland, in the form of a well-structured central statistical office, referred to as Statistics Poland, was possible only thanks to intensive scientific work performed earlier by dozens of Polish statisticians. Their extensive contribution to this process is presented in more detail in the monograph entitled *Polish Statisticians. Biograms* (2018).

REFERENCES

Berger J., (1995), Badania statystyczne na ziemiach polskich do 1918 r., in: S. Jońca (ed.), *Obchody jubileuszowe 200-lecia statystyki polskiej i 75-lecia Głównego Urzędu Statystycznego*, Główny Urząd Statystyczny, Warszawa, 41–44.

Berger J., (2008), Powstanie i pierwsze lata działalności Polskiego Towarzystwa Statystycznego, in: A. Kula (ed.), *Statystyka wczoraj, dziś i jutro*, Główny Urząd Statystyczny, Warszawa, 26–27.

Berger J., (2018), Ludwik Krzywicki (1859–1941), in: M. Krzyśko (ed.), *Statystycy polscy: Biogramy*, Główny Urząd Statystyczny, Warszawa, 212–217.

Domański Cz., (2008), Statystycy zasłużeni dla nauki, in: K. Jakóbik (ed.), *Statystyka społeczna, deklaracje, szanse, perspektywy*, Główny Urząd Statystyczny, Warszawa, 7–22.

Kubiczek F., (ed.) (2017), *Statystyka Polski: Dawniej i dzisiaj*, Główny Urząd Statystyczny, Warszawa.

Kruszka K., (ed.) (2012), *Polskie Towarzystwo Statystyczne 1912–2012*, Polskie Towarzystwo Statystyczne, Warszawa.

Łazowska B., (2015), *Działalność badawcza Głównego Urzędu Statystycznego w okresie II Rzeczypospolitej,* Centralna Biblioteka Statystyczna, Warszawa.

Łazowska B., (2018a), Józef Buzek (1873–1936), in: M. Krzyśko (ed.), *Statystycy polscy: Biogramy*, Główny Urząd Statystyczny, Warszawa, 61–65.

Łazowska B., (2018b), Statystyka na ziemiach polskich pod panowaniem pruskim, *Wiadomości Statystyczne*, (5), 78–102.

Łukasiewicz J., (1995), Polska histografia a statystyka, in: S. Jońca (ed.) *Obchody jubileuszowe 200-lecia statystyki polskiej i 75-lecia Głównego Urzędu Statystycznego*, Główny Urząd Statystyczny, Warszawa, 58–68.

Pociecha J., (2017), Polska statystyka publiczna – od Sejmu Czteroletniego do współczesności, *Folia Oeconomica Cracoviensia*, (58), 5–22.

Romer E., (1916), *Geograficzno-statystyczny atlas Polski*, Gebethner i Wolff, Warszawa.

Roszkowski W., (2010), *Społeczny ruch ekonomiczny w Polsce przed 1939 r.*, Polskie Towarzystwo Ekonomiczne, Warszawa.

PRZEGLĄD STATYSTYCZNY
TOM LXVI – ZESZYT 4 – 2019

Sprawozdanie z XXVIII Konferencji Naukowej „Klasyfikacja i analiza danych – teoria i zastosowania”

Jacek BATÓG[1]
Krzysztof JAJUGA[2]
Marek WALESIAK[3]

XXVIII Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego (SKAD PTS) (XXXIII Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania” odbyła się 18–20 września 2019 r. w Szczecinie. Jej organizatorami były Sekcja Klasyfikacji i Analizy Danych PTS oraz Wydział Ekonomii, Finansów i Zarządzania Uniwersytetu Szczecińskiego (US).

Komitetowi Organizacyjnemu konferencji przewodniczył dr hab. Jacek Batóg, prof. US, a funkcje sekretarzy naukowych pełnili dr hab. Iwona Foryś, prof. US, dr Barbara Batóg, dr Monika Rozkrut i dr Paweł Baran.

Głównymi celami konferencji były: zaprezentowanie osiągnięć i wymiana doświadczeń z zakresu teoretycznych i stosowanych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum umożliwiające badaczom podsumowanie aktualnego stanu wiedzy, przedstawienie i upowszechnianie nowatorskich dokonań oraz wskazanie kierunków dalszych prac i badań.

Tematyka konferencji obejmowała problematykę teoretyczną i praktyczną. Do tej pierwszej należały kwestie taksonomii, analizy dyskryminacyjnej, metod porządkowania liniowego, metod statystycznej analizy wielowymiarowej, metod analizy zmiennych ciągłych, metod analizy zmiennych dyskretnych, metod analizy danych symbolicznych oraz metod graficznych. Wśród dyskutowanych zastosowań znalazły się: analizy danych finansowych, marketingowych, przestrzennych, a także wykorzystanie analizy danych w medycynie, psychologii, archeologii oraz aplikacje komputerowe metod statystycznych.

[1] Uniwersytet Szczeciński, Wydział Ekonomii, Finansów i Zarządzania, Instytut Ekonomii i Finansów, Katedra Ekonometrii i Statystyki, ul. Mickiewicza 64, 71-101 Szczecin, e-mail: jacek.batog@usz.edu.pl, ORCID: https://orcid.org/0000-0003-1413-7692.

[2] Uniwersytet Ekonomiczny we Wrocławiu, Wydział Ekonomii i Finansów, Katedra Inwestycji Finansowych i Zarządzania Ryzykiem, ul. Komandorska 118-120, 53-345 Wroclaw, e-mail: krzysztof.jajuga@ue.wroc.pl, ORCID: https://orcid.org/0000-0002-5624-6929.

[3] Uniwersytet Ekonomiczny we Wrocławiu, Wydział Ekonomii i Finansów, Katedra Ekonometrii i Informatyki, ul. Nowowiejska 3, 58-500 Jelenia Góra, e-mail: marek.walesiak@ue.wroc.pl, ORCID: https://orcid.org/0000-0003-0922-2323.

W konferencji wzięło udział 85 pracowników oraz doktorantów reprezentujących: Akademię Górniczo-Hutniczą w Krakowie, Akademię im. Jakuba z Paradyża w Gorzowie Wlkp., Akademię Pomorską w Słupsku, Politechnikę Opolską, Państwową Wyższą Szkołę Zawodową w Skierniewicach, Szkołę Główną Gospodarstwa Wiejskiego w Warszawie, Szkołę Główną Handlową w Warszawie, Uniwersytet Ekonomiczny w Katowicach, Uniwersytet Ekonomiczny w Krakowie, Uniwersytet Ekonomiczny w Poznaniu, Uniwersytet Ekonomiczny we Wrocławiu, Uniwersytet Gdański, Uniwersytet Łódzki, Uniwersytet im. Adama Mickiewicza w Poznaniu, Uniwersytet Medyczny im. Karola Marcinkowskiego w Poznaniu, Uniwersytet Mikołaja Kopernika w Toruniu, Uniwersytet Przyrodniczy w Poznaniu, Uniwersytet Szczeciński, Uniwersytet Rolniczy w Krakowie, Zachodniopomorski Uniwersytet Technologiczny w Szczecinie, a także Główny Urząd Statystyczny i urzędy statystyczne w Szczecinie, Poznaniu i we Wrocławiu.

W trakcie dwóch sesji plenarnych oraz dwunastu sesji równoległych wygłoszono 44 referaty. Odbyła się również sesja plakatowa, licząca 32 wystąpienia. Poszczególnym sesjom przewodniczyli: Andrzej Sokołowski, Krzysztof Jajuga, Eugeniusz Gatnar, Iwona Konarzewska, Marek Walesiak, Mirosław Krzyśko, Grażyna Dehnel, Barbara Pawełek, Józef Pociecha, Andrzej Bąk, Grażyna Trzpiot, Feliks Wysocki, Andrzej Dudek, Krzysztof Najman.

Wygłoszono następujące referaty:

- Krzysztof Jajuga, *Data Science a analiza danych – nowa jakość czy nowe szaty?*;
- Eugeniusz Gatnar, *Modele pomiaru oczekiwań inflacyjnych w Polsce*;
- Barbara Pawełek i Józef Pociecha, *Problem wartości odstających w prognozowaniu upadłości przedsiębiorstw z wykorzystaniem metody Logit leaf model;*
- Paweł Ulman i Małgorzata Ćwiek, *Ubóstwo mieszkaniowe polskich gospodarstw domowych oraz jego zróżnicowanie*;
- Iwona Bąk i Katarzyna Wawrzyniak, *Wykorzystanie metod wielowymiarowej analizy statystycznej w badaniu wypalenia zawodowego nauczycieli i wykładowców wyższych uczelni*;
- Joanna Landmesser, *Różnice w rozkładach dochodów kobiet i mężczyzn w krajach Unii Europejskiej*;
- Agnieszka Wałęga i Grzegorz Wałęga, *Psychologiczne konsekwencje nadmiernego zadłużenia gospodarstw domowych w Polsce*;
- Marek Walesiak i Grażyna Dehnel, *Zastosowanie skalowania wielowymiarowego w ocenie zmian w procesie starzenia się ludności województw Polski dla lat 2002, 2010 i 2017*;
- Jacek i Barbara Batógowie, *Analiza wpływu selekcji zmiennych na zgodność i stabilność klasyfikacji*;
- Kamila Migdał-Najman i Krzysztof Najman, *Algorytmy przyrostowe w analizie skupień*;

- Mariusz Kubus, *Wpływ niezbilansowania próby uczącej na dobór zmiennych i jakość klasyfikatorów*;
- Romana Głowicka-Wołoszyn i Feliks Wysocki, *Problem identyfikacji stopnia samodzielności finansowej gmin ze względu na występowanie asymetrii rozkładów cech prostych*;
- Przemysław Szczuciński, *Badanie atrakcyjności turystycznej gmin województwa lubuskiego z wykorzystaniem miary agregatowej*;
- Romana Głowicka-Wołoszyn, *Koncepcja przestrzennego taksonomicznego miernika potencjału. Przykład zastosowania w ocenie samodzielności finansowej gmin województwa wielkopolskiego*;
- Adam Sagan i Mariusz Grabowski, *Analiza wrażliwości w identyfikacji przyczynowych efektów mediacji w modelu TAM*;
- Justyna Majewska i Grażyna Trzpiot, *Modelowanie wskaźnika umieralności dla jednej i dwóch populacji dla wybranych krajów Europy Środkowej i Wschodniej*;
- Marcin Pełka i Aneta Rybicka, *Identyfikacja czynników decydujących o odejściu klienta z zastosowaniem hybrydowego modelu dla danych symbolicznych*;
- Paweł Baranowski, Karol Korczak i Jarosław Zając, *Prognozowanie frekwencji na seansach kinowych w Polsce*;
- Maria Kaźmierska-Zatoń i Wojciech Zatoń, *Rola barier działalności przedsiębiorstw przemysłowych w ocenie koniunktury gospodarczej w Polsce*;
- Barbara Batóg i Katarzyna Wawrzyniak, *Porównanie propozycji transformacji nominant na stymulanty na przykładzie wskaźników finansowych dla spółek notowanych na GPW*;
- Dominik Krężołek, *Zastosowanie estymatora Hilla do oceny ryzyka ekstremalnego na rynku metali*;
- Katarzyna Cheb i Iwona Bąk, *Wykorzystanie wielowymiarowej analizy statystycznej w identyfikacji powiązań pomiędzy zrównoważonym rozwojem i zieloną gospodarką w krajach Unii Europejskiej*;
- Mariusz Doszyń, *Ekonometryczny model cen nieruchomości z restrykcjami w postaci nierówności*;
- Katarzyna Frodyma, Monika Papież i Sławomir Śmiech, *Oddzielenie wzrostu gospodarczego od emisji dwutlenku węgla w krajach Unii Europejskiej*;
- Dorota Rozmus, *Miejsce Polski w UE pod względem zrównoważonego rozwoju w świetle różnych miar stabilności grupowania*;
- Iwona Markowicz i Paweł Baran, *Przegląd wskaźników jakości danych dotyczących wewnątrzwspólnotowych dostaw towarów. Wybór wskaźnika a uporządkowanie krajów*;
- Marcin Salamaga, *Propozycja badania zależności pomiędzy konkurencyjnością technologiczną i eksportową krajów Europy Środkowo-Wschodniej*;

- Iwona Konarzewska, *Wpływ siły i struktury współzależności liniowej kryteriów na wyniki wielokryteriowej analizy porównawczej*;
- Andrzej Dudek, *Indeks sylwetkowy jako narzędzie oceny jakości klasyfikacji*;
- Beata Bieszk-Stolorz, *Modele Prentice'a-Williamsa-Petersona w ocenie wpływu cech osób bezrobotnych na intensywność kolejnych zarejestrowań w urzędzie pracy*;
- Wojciech Roszka i Marcin Szymkowiak, *Estymacja wybranych charakterystyk rynku pracy w ujęciu przestrzennym z wykorzystaniem techniki symulowanego wyżarzania*;
- Andrzej Sokołowski i Małgorzata Markowska, *Elastyczne porządkowanie liniowe obiektów*;
- Jacek Białek, *Pomiar inflacji przy zastosowaniu danych skanowanych*;
- Bartłomiej Jefmański, *Syntetyczna miara rozwoju Hellwiga dla intuicjonistycznych zbiorów rozmytych*;
- Grzegorz Szczerbak, *Analiza techniczna z wykorzystaniem algorytmów rozpoznawania obrazów*;
- Grażyna Trzpiot i Justyna Majewska, *Wpływ długowieczności na wycenę zwrotów z inwestycji długoterminowych: przypadek wybranych krajów europejskich*;
- Andrzej Bąk, *Pomiar gotowości do zapłaty za dobra nierynkowe z wykorzystaniem analizy conjoint*;
- Adam Idczak, *Porównanie wybranych metod analizy text miningowej do klasyfikacji tekstów marketingowych*;
- Paweł Baran, *Wykorzystanie technik word embedding do poprawy jakości baz danych handlu zagranicznego – studium empiryczne*;
- Tomasz Górecki, Mirosław Krzyśko i Waldemar Wołyński, *Mierzenie i testowanie wzajemnej zależności dla danych funkcjonalnych*;
- Artur Zaborski, *Zastosowanie niepełnej metody tetrad do pomiaru podobieństw w niemetrycznym skalowaniu wielowymiarowym*;
- Marcin Pełka, *Wzmocnienie zdolności predykcyjnych podejścia wielomodelowego danych symbolicznych z zastosowaniem wyodrębniania zmiennych objaśniających z sieci neuronowych*;
- Wioletta Grzenda, *Dyskretyzacja zmiennych ciągłych a moc predykcyjna klasyfikatorów na przykładzie modeli regresji logistycznej oraz sztucznych sieci neuronowych*;
- Dominik Rozkrut, *AI i ML w statystyce publicznej: stan i perspektywy rozwoju*.

Pierwszego dnia konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych PTS pod przewodnictwem Krzysztofa Jajugi. Plan zebrania obejmował następujące punkty:

- sprawozdanie z działalności SKAD PTS;
- informacje dotyczące planowanych konferencji krajowych i zagranicznych;

- organizacja konferencji SKAD PTS w następnych latach;
- wybór przedstawiciela SKAD PTS do International Federation of Classification Societies (IFCS) Council na kadencję 2020–2023.

Sprawozdanie z działalności SKAD PTS przedstawiła Barbara Pawełek, sekretarz naukowy sekcji. Poinformowała, że sekcja liczy obecnie 236 członków, a także przypomniała, że na stronie internetowej sekcji znajduje się regulamin oraz deklaracja członkowska.

Kolejna informacja dotyczyła artykułów po konferencji SKAD 2018 (10––12 września 2018 r. w Ciechocinku), które można było składać do czasopism: „Statistics in Transition new series" (w języku angielskim), „Przegląd Statystyczny" (w języku polskim lub angielskim), „Econometrics. Advances in Applied Data Analysis" (w języku angielskim), „Dynamic Econometric Models" (w języku angielskim) oraz „Folia Oeconomica Stetinensia" (również po angielsku). W zeszycie 1 „Przeglądu Statystycznego" z 2019 r. ukazało się sprawozdanie z konferencji autorstwa Krzysztofa Jajugi, Tadeusza Kufla i Marka Walesiaka (s. 94–99).

Następnie Barbara Pawełek przedstawiła informacje dotyczące działalności międzynarodowej oraz udziału członków SKAD w ważnych konferencjach:

- GPSDAA 2019, Bayreuth, Niemcy, 17 marca 2019 r. (Komitet Naukowy: Józef Pociecha, przewodniczenie sesji: Józef Pociecha, Józef Dziechciarz, wystąpienia z Polski: 7);
- ECDA 2019, Bayreuth, Niemcy, 18–20 marca 2019 r. (Komitet Naukowy: Krzysztof Jajuga, Józef Pociecha, Adam Sagan, przewodniczenie sesji: Krzysztof Jajuga (dwukrotnie), Józef Pociecha, Adam Sagan, Józef Dziechciarz, wystąpienia z Polski: 13);
- IFCS 2019, Saloniki, Grecja, 26–29 sierpnia 2019 r. (Komitet Naukowy: Krzysztof Jajuga, przewodniczenie sesji: Krzysztof Jajuga, Józef Pociecha, Tomasz Górecki, wystąpienia z Polski: 10);
- CLADAG 2019, Cassino, Włochy, 11–13 września 2019 r. (Komitet Naukowy: Józef Pociecha, przewodniczenie sesji: Józef Pociecha, wystąpienia z Polski: 2).

Kolejny punkt posiedzenia sekcji obejmował zapowiedzi pokrewnych tematycznie konferencji krajowych: XXXVIII Konferencji Naukowej „Multivariate Statistical Analysis MSA 2019" (Łódź, 4–6 listopada 2019 r.), XIV Międzynarodowej Konferencji Naukowej im. Profesora Aleksandra Zeliasia „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych" (Zakopane, 12–14 maja 2020 r.) oraz II Konferencji Naukowej z cyklu Metodologia Badań Statystycznych (MET2020, Warszawa, 1–3 lipca 2020 r.), jak również konferencji zagranicznych: ECDA 2020 (Neapol, 16–18 września 2020 r.), IFCS 2021, CLADAG 2021 i GPSDAA 2021. W trakcie posiedzenia podjęto też kwestię organizacji kolejnych konferencji SKAD. We wrześniu 2020 r. konferencję zorganizuje w Sopocie Uniwersytet Gdański.

W dalszej części zebrania dokonano wyboru przedstawiciela SKAD do IFCS Council. Powołano komisję skrutacyjną w składzie: Iwona Konarzewska – przewodnicząca, Andrzej Bąk – członek i Jacek Batóg – członek. Krzysztof Jajuga zgłosił kandydaturę Andrzeja Sokołowskiego, który wyraził zgodę na kandydowanie. Następnie zgłoszony i jednogłośnie poparty został wniosek o zamknięcie listy kandydatów. W tajnym głosowaniu uczestniczyło 27 członków sekcji, jednogłośnie popierając kandydaturę Andrzeja Sokołowskiego.

Na zakończenie konferencji ogłoszono przyznanie nagrody firmy StatSoft na najlepszy plakat. Jury w składzie Józef Pociecha (przewodniczący), Mirosław Krzyśko i Iwona Konarzewska zadecydowało o przyznaniu nagrody Małgorzacie Misztal i Aleksandrze Kupis-Fijałkowskiej za plakat *Rozwój technologii informacyjno-telekomunikacyjnych w Polsce na tle Unii Europejskiej – wielowymiarowa analiza statystyczna*.

Teksty referatów konferencyjnych, które zostaną przygotowane w formie recenzowanych artykułów naukowych, ukażą się w monografii *Classification and Data Analysis – Theory and Applications, Springer: Studies in Classification, Data Analysis, and Knowledge Organization* pod red. Jajugi, Batóga i Walesiaka, która ukaże się nakładem wydawnictwa Springer. Ponadto artykuły pokonferencyjne publikowane będą w czasopismach „Folia Oeconomica Stetinensia", „Econometrics. Advances in Applied Data Analysis" i „Przegląd Statystyczny".

Informacje o konferencji dostępne są na stronie www.skad2019.usz.edu.pl.

Podziękowania dla recenzentów

Redakcja *Przeglądu Statystycznego* składa podziękowania Recenzentom artykułów nadesłanych do publikacji w 2019 roku:

Tomasz Bartłomowicz
Andrzej Bąk
Michał Bernardelli
Justyna Brzezińska
Andrius Čiginas
Mikołaj Czajkowski
Grażyna Dehnel
Andrzej Dudek
Józef Dziechciarz
Waldemar Florczak
Elżbieta Gołata
Joanna Górka
Wojciech Grabowski
Michał Gradzewicz
Paweł Hanczar
Beata Jackowska
Jerzy Jakubczyc
Anna Janicka
Bartłomiej Jefmański
Katarzyna Kopczewska
Maciej Kostrzewski
Marek Kośny
Barbara Kowalczyk
Mariusz Kubus
Ireneusz Kuropka
Mariusz Łapczyński
Bożena Łazowska
Kamila Migdał-Najman
Andrzej Młodak
Joanicjusz Nazarko
Arkadiusz Orzechowski
Walenty Ostasiewicz
Tomasz Panek
Barbara Pawełek
Krzysztof Piasecki
Michał Pietrzak
Józef Pociecha
Artur Prędki
Ewa Putek-Szeląg
Ewa Roszkowska
Alet Roux
Marcin Salamaga
Krzysztof Targiel
Tomasz Tokarski
Grażyna Trzpiot
Joanna Wolszczak-Derlacz
Feliks Wysocki
Henryk Zawadzki
Ren Zhang