Wiadomości Statystyczne. The Polish Statistician, 2022, vol. 67, 12, 62–79 DOI: 10.5604/01.3001.0016.1615
Statystyka w praktyce / Statistics in practice
Zgłoszony/submitted: 21.06.2022, poprawiony/revised: 9.11.2022, zaakceptowany/accepted: 18.11.2022

Dylematy związane z estymacją dominanty wynagrodzeń

Mirosław Błażej[a], Emilia Gosińska[b]

Streszczenie. Dominanta wynagrodzeń to ważny wskaźnik opisujący rozkład wynagrodzeń, ale ze względu na silną asymetrię rozkładu tej cechy w Polsce jej wyznaczenie nie należy do standardowych działań w analizie struktury wynagrodzeń. Celem artykułu jest omówienie wybranych metod szacowania dominanty oraz porównanie wyników jej estymacji otrzymanych za pomocą różnych metod. Wykorzystano dane o wynagrodzeniach indywidualnych brutto w październiku 2018 r. pochodzące z badania struktury wynagrodzeń przeprowadzonego przez Główny Urząd Statystyczny.

W przypadku metody standardowej, wykorzystującej wzór interpolacyjny i histogram, wartość oszacowanej dominanty jest wrażliwa na założoną rozpiętość przedziałów w szeregu rozdzielczym i początek pierwszego przedziału. Zmniejszanie rozpiętości przedziałów powoduje dążenie dominanty do wartości równej płacy minimalnej. Zastosowanie zaawansowanych metod statystycznych, m.in. wykorzystujących estymator jądrowy, prowadzi do otrzymania znacząco różnych oszacowań dominanty w zależności od metody (rozrzut wyników wynosi ok. 800 zł).

Analiza otrzymanych wyników daje ponadto podstawy do rozważenia tezy, że rozkład wynagrodzeń jest mieszany: ma cechy rozkładu dyskretnego dla wynagrodzeń w wysokości płacy minimalnej i ciągłego – dla wynagrodzeń powyżej płacy minimalnej oraz odznacza się cyklicznością (w Polsce zawiera się więcej umów, w których kwota wynagrodzenia jest wielokrotnością 50 zł lub 100 zł, niż umów na inne kwoty).

Słowa kluczowe: dominanta, estymator jądrowy, estymacja dominanty, histogram

JEL: C46

Dilemmas relating to mode estimation of wages and salaries

Abstract. The mode of wages and salaries is an important indicator describing their distribution; however, due to the strong asymmetry of the distribution of this feature in Poland, mode estimation is not a standard procedure in the analysis of the structure of wages and salaries. The aim of the article is to discuss selected methods of estimating the mode and to compare the mode estimation results obtained by means of various methods. The research is

[a] Główny Urząd Statystyczny, Departament Studiów Makroekonomicznych i Finansów, Polska / Statistics Poland, Department of Macroeconomic Studies and Finance, Poland.
ORCID: https://orcid.org/0000-0003-4482-8996. E-mail: m.blazej@stat.gov.pl.

[b] Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny; Główny Urząd Statystyczny, Departament Studiów Makroekonomicznych i Finansów; Polska / University of Lodz, Faculty of Economics and Sociology; Statistics Poland, Department of Macroeconomic Studies and Finance; Poland.
ORCID: https://orcid.org/0000-0002-5325-6144. Autor korespondencyjny / Corresponding author, e-mail: emilia.gosinska@uni.lodz.pl.

based on data on individual gross wages and salaries registered in October 2018 in Poland. The data came from a survey of the structure of wages and salaries conducted by Statistics Poland.

In the case of the standard method based on the interpolation formula and histogram, the mode estimate is sensitive to the assumed span of intervals in the frequency table and the beginning of the first interval. Reducing the span of the intervals causes the mode to reach the value of the minimum wage. The application of advanced methods, including those using a kernel estimator, leads to significantly different estimates of the mode depending on the method used (the dispersion reaches the value of approximately PLN 800).

Additionally, the analysis of the obtained results gives grounds to considering a thesis that wage and salary distribution is a mixture of the following distributions: discrete (for the minimum wage) and continuous (for wages and salaries above the minimum wage), and is characterised by cyclicality (in Poland, more contracts offer remunerations which are a multiple of PLN 50 or PLN 100 than remunerations for other amounts).

Keywords: mode, kernel estimator, mode estimation, histogram

1. Wprowadzenie

Dominanta (moda, modalna, wartość najczęstsza, wartość najbardziej prawdopodobna) jest jedną z miar tendencji centralnej, należącą do grupy miar pozycyjnych. Sposób jej obliczania oraz definicja zależą od typu rozpatrywanych zmiennych. Wyznaczenie dominanty w szeregach szczegółowych i rozdzielczych punktowych polega na wskazaniu wartości cechy, której odpowiada największa liczebność (Parlińska i Parliński, 2011). W szeregach rozdzielczych z przedziałami klasowymi można określić przedział, w którym występuje modalna, a jej przybliżoną wartość wyznacza się za pomocą interpolacji.

Obliczenie dominanty dla zmiennych ciągłych oraz quasi-ciągłych jest zagadnieniem nietrywialnym, ponieważ wymaga przybliżenia rozkładu badanej zmiennej (w przypadku zmiennej ciągłej dwie wartości pobrane z tego samego rozkładu nie są sobie równe, a to uniemożliwia znalezienie wartości występującej najczęściej). Najprostszym i najpopularniejszym sposobem przybliżenia rozkładu zmiennej jest histogram. Bardziej zaawansowane metody polegają na estymacji rozkładu funkcji gęstości prawdopodobieństwa z wykorzystaniem np. estymatorów jądrowych (Parzen, 1962). Wówczas modalna jest wartością, która odpowiada maksimum krzywej najlepiej dopasowanej do rozkładu rzeczywistego (maksimum oszacowanej funkcji gęstości).

W artykule rozważane są dwa rodzaje metod estymacji dominanty. Pierwszy z nich to najprostsza dwukrokowa procedura (dalej zwana *metodą standardową*), w której najpierw należy skonstruować szereg rozdzielczy i znaleźć przedział z największą liczbą przypadków, a następnie zastosować wzór interpolacyjny (Ostasiewicz i in., 1995; Stanisławek, 2010). Drugi rodzaj metod estymacji dominanty to zaawansowane metody nieparametryczne (dalej zwane także *zaawansowanymi metodami statystycznymi*), wykorzystujące np. algorytmy iteracyjne, numeryczną optymalizację funkcji gęstości, a także estymatory jądrowe.

Zaawansowane metody szacowania dominanty obejmują zarówno nieparametryczne (m.in. Fukunaga i Hostetler, 1975; Tsybakov, 1990), jak i parametryczne metody estymacji funkcji gęstości (Bickel, 2002, 2003), a także wykorzystują złożone podejścia numeryczne (Bickel i Frühwirth, 2006; Grenander, 1965; Lientz, 1972; Venter, 1967). Jurkiewicz i Kozłowski (2009) oraz Sokołowski (2013) porównali efektywność wybranych estymatorów dominanty, wykorzystując metody symulacyjne.

W badaniu omawianym w niniejszym artykule porównano wyniki estymacji dominanty metodą standardową i zaawansowanymi metodami statystycznymi. Posłużono się danymi o wynagrodzeniach brutto w Polsce pochodzącymi z reprezentacyjnego badania struktury wynagrodzeń według zawodów za październik 2018 r., przeprowadzonego przez Główny Urząd Statystyczny (GUS) w przedsiębiorstwach o liczbie pracujących powyżej dziewięciu osób[1]. Przy założeniu, że rozkład wynagrodzeń jest dyskretny, największa liczba wystąpień przypada dla wartości wynagrodzenia minimalnego. Wartość ta reprezentuje jednak małą frakcję przypadków w relacji do liczebności badanej zbiorowości. Z uwagi na to celem artykułu jest eksploracja wybranych metod szacowania dominanty dla danych quasi-ciągłych, porównanie wyników estymacji dominanty różnymi metodami na podstawie danych dotyczących wynagrodzeń indywidualnych brutto w Polsce za październik 2018 r. i sprawdzenie, czy pozwolą one na oszacowanie dominanty wynagrodzeń reprezentującej większą frakcję przypadków, która ma uzasadnienie metodologiczne i społeczno-gospodarcze oraz odzwierciedla sytuację na rynku pracy. W badaniu podjęto próbę wskazania przedziału wartości dominanty, do którego należałyby jej wartości obliczone ze standardowego wzoru interpolacyjnego przy założeniu różnej szerokości przedziałów szeregu rozdzielczego, a także z wykorzystaniem bardziej zaawansowanych estymatorów modalnej jednowymiarowej dostępnych w pakiecie modeest w programie R (Poncet, 2019).

2. Metody estymacji dominanty

W artykule opisano dwa rodzaje metod estymacji dominanty: standardową metodę dwukrokową oraz metodę numerycznej optymalizacji funkcji gęstości z wykorzystaniem estymacji jądrowej jako przykład zaawansowanych metod estymacji dominanty.

2.1. Metoda standardowa

Właściwym sposobem prezentacji danych dotyczących wynagrodzeń jest zbudowanie szeregu rozdzielczego z przedziałami. Uzyskuje się podział zbiorowości statystycznej na klasy według wartości określonej cechy, a każdej z wyodrębnionych klas

[1] Szczegółowe informacje dotyczące badania są zawarte w publikacjach: *Struktura wynagrodzeń według zawodów w październiku 2018 r.* (GUS, 2020b) oraz *Zeszyt metodologiczny. Struktura wynagrodzeń według zawodów* (GUS, 2020c).

przyporządkowana zostaje odpowiednia liczba jednostek (Ostasiewicz i in., 1995). W szeregach rozdzielczych (przedziałowych) można bezpośrednio wyznaczyć przedział, w którym występuje dominanta, i jest to przedział z największą liczbą wartości do niego należących. Powszechnie stosowaną formę graficznej prezentacji takiego typu szeregu stanowi histogram częstości, na podstawie którego można wskazać przedział o największej liczebności. W przypadku danych pogrupowanych w szereg rozdzielczy z przedziałami nie wystarczy znalezienie przedziału o największej liczebności; należy także wskazać punkt dominantowy – konkretny punkt w tym przedziale (Stanisławek, 2010). Przybliżoną wartość dominanty dla szeregu rozdzielczego wyznacza się graficznie z histogramu liczebności lub ze wzoru interpolacyjnego (Ostasiewicz i in., 1995; Sokołowski, 2013; Stanisławek, 2010):

$$D_0 = x_0 + \frac{(n_0 - n_{-1}) \cdot h_0}{(n_0 - n_{-1}) + (n_0 - n_{+1})}, \tag{1}$$

gdzie:

x_0 – dolna granica przedziału klasowego, który zawiera dominantę (przedział charakteryzuje się największą liczebnością/częstością),

n_0 – liczebność/częstość (absolutna lub względna) przedziału klasowego zawierającego dominantę,

n_{-1} – liczebność/częstość (absolutna lub względna) przedziału klasowego poprzedzającego przedział zawierający dominantę,

n_{+1} – liczebność/częstość (absolutna lub względna) przedziału klasowego następującego po przedziale zawierającym dominantę,

h_0 – rozpiętość przedziału dominanty (przedziały sąsiadujące muszą mieć taką samą rozpiętość).

Wyznaczenie dominanty za pomocą wzoru (1) jest możliwe wtedy, gdy spełnione są następujące warunki (GUS, 2020a; Ostasiewicz i in., 1995):

- występuje wystarczająco dużo obserwacji;
- rozkład empiryczny liczebności jest rozkładem jednomodalnym (występuje jedno wyraźnie zaznaczone maksimum);
- asymetria rozkładu liczebności jest umiarkowana (dominanta nie występuje w skrajnym przedziale);
- wszystkie przedziały są tej samej długości.

W przypadku danych pogrupowanych w szereg rozdzielczy rozpiętość przedziału dominanty ściśle zależy od przyjętej liczby przedziałów. Dla różnych rozpiętości przedziałów w szeregu rozdzielczym dominanta wyznaczona ze wzoru (1) przyjmuje różne wartości i wybranie tej prawidłowej jest arbitralne. Przy założeniu malejącej długości przedziałów pewnym oszacowaniem modalnej mogłaby być wartość, do której zbiegałyby wartości dominanty wyznaczone dla różnych szerokości przedziałów.

Inną metodą prezentacji rozkładu cechy jest rozkład decylowy (GUS, 2020b, tabl. 13, przedstawiająca najwyższe miesięczne wynagrodzenia w grupach decylowych). W przypadku rozkładu decylowego pewnym przybliżeniem dominanty jest przedział o najmniejszej rozpiętości.

2.2. Zaawansowane metody statystyczne

Alternatywną dla histogramu i jednocześnie bardziej zaawansowaną numerycznie nieparametryczną metodą wyznaczenia dominanty dla cech quasi-ciągłych oraz prób charakteryzujących się dużą liczebnością jest metoda estymacji funkcji gęstości prawdopodobieństwa i dominanty zaproponowana przez Parzena (1962). Oszacowanie funkcji gęstości rozkładu można zdefiniować, wykorzystując estymator jądrowy gęstości (ang. *kernel density estimation*) w następujący sposób:

$$f_n(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right), \tag{2}$$

gdzie:
$X_1, X_2, \ldots, X_n$ – niezależne zmienne losowe o identycznym rozkładzie,
h – odpowiednio dobrana liczba dodatnia,
$K(y)$ – funkcja zdefiniowana następująco:

$$K(y) = \begin{cases} \frac{1}{2} \text{ dla } |y| \leq 1 \\ 0 \text{ dla } |y| > 1 \end{cases}.$$

Istnieje wiele możliwych oszacowań $f_n(x)$ funkcji gęstości rozkładu prawdopodobieństwa $f(x)$ w zależności od założonego parametru h i postaci funkcji $K(\cdot)$. Jeśli $K(\cdot)$ spełnia odpowiednie założenia (zob. Parzen, 1962), jest nazywana funkcją ważącą (ang. *weighting function*).

W pracy Parzena (1962) omówione zostały przykłady funkcji $K(\cdot)$ oraz założenia, jakie należy spełnić, aby oszacowana funkcja gęstości prawdopodobieństwa $f_n(x)$ była jednostajnie zbieżna do prawdziwej funkcji gęstości. Jeśli są one spełnione, rozważany estymator jest jednostajnie zgodny i istnieje zgodne oszacowanie dominanty z próby mo_n:

$$f_n(mo_n) = max_{-\infty<x<\infty} f_n(x), \tag{3}$$

gdzie n oznacza liczebność próby.

W tej samej pracy udowodniono także asymptotyczną normalność estymatora dominanty z próby przy założeniu, że są spełnione określone warunki. Estymator jądrowy umożliwia oszacowanie gęstości dowolnego rozkładu (w tym rozkładów nietypowych, wielomodalnych[2]). Jeżeli rozkład badanej zmiennej jest wielomodalny, to wyznaczenie lokalnych maksimów pozwala wskazać dominanty lokalne. Właściwe oszacowanie dominanty za pomocą estymatorów jądrowych jest wrażliwe na zastosowanie funkcji $K(\cdot)$, które odwzorują rozkład teoretyczny badanej zmiennej.

Oprócz metody Parzena można wskazać inne metody nieparametryczne wykorzystujące estymator jądrowy, m.in. estymator Tsybakova (Tsybakov, 1990) oraz estymator Meanshift (Fukunaga i Hostetler, 1975). Do metod nieparametrycznych, które wykorzystują zaawansowane metody numeryczne, ale nie uwzględniają estymacji funkcji gęstości, można zaliczyć: estymator Grenandera (Grenander, 1965), estymator Lientza (minimalizacja funkcji Lientza; Lientz, 1972), estymator Ventera (Venter, 1967), metodę półprób HSM (ang. *half-sample mode*; Bickel i Frühwirth, 2006) czy metodę półrozstępów HRM (ang. *half-range mode*; Bickel, 2002). Natomiast parametryczne metody estymacji to m.in. standardowa metoda parametryczna (ang. *standard parametric mode*) oraz odporna metoda parametryczna (ang. *robust parametric mode*; Bickel, 2003).

2.3. Porównanie metod

Dobór metody estymacji dominanty zależy od specyfiki danych statystycznych. Zaletą dominanty jest jej odporność na skrajne wartości rozkładu, jednak w przypadku rozkładów skrajnie asymetrycznych efektywność metody standardowej wykorzystującej wzór interpolacyjny maleje. Na podstawie wyników symulacji Jurkiewicz i Kozłowski (2009) zauważają, że im większa skośność, tym więcej przedziałów jest wymaganych do efektywnego szacunku.

Metody parametryczne dostarczają efektywniejszych ocen rzeczywistej dominanty (zob. symulacje w pracy Jurkiewicza i Kozłowskiego, 2009). Wadą tych metod jest jednak ich złożoność obliczeniowa – konieczne staje się iteracyjne ustalenie parametru najlepiej dopasowanej funkcji potęgowej. Zaawansowane nieparametryczne metody estymacji dominanty są wrażliwe na zastosowanie odpowiedniej postaci funkcji (zob. wzór (2)) oraz innych parametrów specyficznych dla danej metody.

W zestawieniu estymatorów modalnej jednowymiarowej sporządzonym na podstawie wyników symulacji w pracy Sokołowskiego (2013) najwyższe pozycje zajęły estymatory: Chernoffa (metoda naiwna; Poncet, 2019), HSM oraz HRM. Należy jednak wziąć pod uwagę to, że modele generowane w omawianych symulacjach nie

[2] W praktyce można wyróżnić rozkłady jednomodalne (w których jest tylko jedna moda), wielomodalne (występuje więcej niż jedna moda) oraz takie, w których moda nie istnieje.

miały cech rozkładów skrajnie asymetrycznych, do których należy rozkład wynagrodzeń analizowany w badaniu omawianym w niniejszym artykule.

3. Wyniki badania

Analiza empiryczna polegała na estymacji dominanty za pomocą metody standardowej wykorzystującej wzór interpolacyjny (1) i wybranych zaawansowanych metod statystycznych. Wykorzystano dane o wynagrodzeniach brutto pochodzące z reprezentacyjnego badania struktury wynagrodzeń według zawodów za październik 2018 r. przeprowadzonego przez GUS w przedsiębiorstwach o liczbie pracujących powyżej dziewięciu osób (GUS, 2020b). Dane te są cechami quasi-ciągłymi charakteryzującymi się dużą liczebnością.

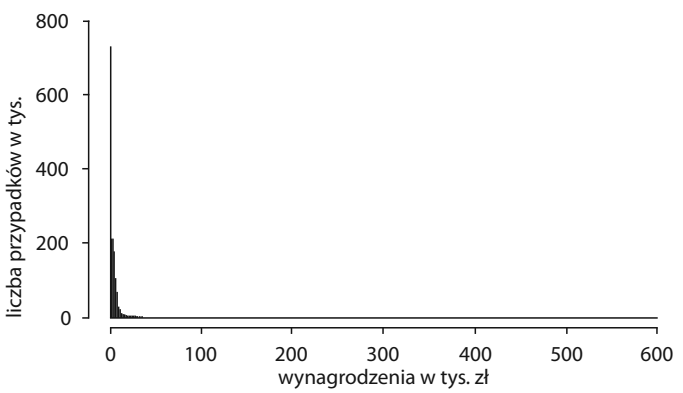
W tabl. 1 przedstawiono wybrane charakterystyki analizowanej populacji pracowników. Histogram dla tej populacji (wykr. 1), liczącej 8425,2 tys. zatrudnionych, świadczy o silnej asymetrii rozkładu wynagrodzeń.

Tabl. 1. Wybrane miary tendencji centralnej charakteryzujące populację zatrudnionych

Miary tendencji centralnej	Wynagrodzenia indywidualne brutto w zł
Wartość minimalna	1763
Kwartyl pierwszy	2913
Mediana	4095
Średnia	5004
Kwartyl trzeci	5718
Wartość maksymalna	598606

Źródło: obliczenia własne wykonane w programie R na podstawie: GUS (2020b).

Wykr. 1. Histogram dla całej populacji zatrudnionych



Źródło: obliczenia własne z wykorzystaniem programu R na podstawie: GUS (2020b).

3.1. Estymacja dominanty metodą standardową

Na pierwszym etapie badania oszacowano wartość dominanty metodą standardową wykorzystującą wzór (1). W tym celu pogrupowano obserwacje w szereg rozdzielczy o zadanej rozpiętości przedziałów. Najpierw przyjęto rozpiętość przedziałów na poziomie 10% średniej wartości wynagrodzenia w populacji, czyli 500,37 zł; dodatkowo założono, że początek pierwszego przedziału w szeregu rozdzielczym wynosi 0. Na podstawie tych założeń obliczono liczebność w przedziale dominanty i przedziałach sąsiednich (tabl. 2, wiersze 1–3). W celu zbadania, czy wartość dominanty wyznaczonej ze wzoru (1) jest wrażliwa na początek pierwszego przedziału, analogicznie obliczono liczebność przy założeniu, że dolna granica pierwszego przedziału jest równa minimalnej wartości cechy, czyli 1763 zł.

Tabl. 2. Liczebność przedziału dominanty i przedziałów sąsiednich przy rozpiętości przedziałów równej 500,37 zł

Przedziały	Granica przedziału		Rozpiętość przedziału	Liczba przypadków	Oszacowana dominanta
	dolna (od)	górna (do)			
	w zł				
Początek pierwszego przedziału równy 0					
Poprzedzający	1501,15	2001,52	500,37	3810	.
Dominanty	2001,53	2501,90	500,37	1 361 267	2379,67
Następujący	2501,91	3002,28	500,37	922481	.
Początek pierwszego przedziału równy 1763					
Pierwszy	1763,01	2263,38	500,37	913 427	.
Poprzedzający	2263,39	2763,76	500,37	900 956	.
Dominanty	2763,77	3264,14	500,37	935 996	3013,79
Następujący	3264,15	3764,52	500,37	900 911	.

Uwaga. Przedział dominanty oznaczono pogrubieniem.

Źródło: obliczenia własne wykonane w programie R na podstawie: GUS (2020b).

Na podstawie danych z tabl. 2 można sformułować następujące wnioski:

- w przypadku przyjęcia 0 jako początku pierwszego przedziału (podczas gdy minimalna wartość cechy wynosi 1763,20 zł), pierwszy przedział ma w rzeczywistości mniejszą rozpiętość niż pozostałe, ponieważ 2001,52 zł – 1763,20 zł = 238,32 zł, a z uwagi na minimalną wartość wynagrodzenia w badanej zbiorowości (1763,20 zł) w przedziale od 1501,15 zł do 1763,20 zł nie ma żadnych przypadków badanej cechy;
- jeśli rzeczywista rozpiętość pierwszego przedziału różni się od pozostałych, wówczas przedziały sąsiadujące z przedziałem dominanty mają różną rozpiętość i nie jest spełniony warunek konieczny do zastosowania wzoru (1);
- w przypadku rozpiętości przedziałów wynoszącej 500,37 zł założenie dotyczące dolnej granicy pierwszego przedziału ma istotny wpływ na oszacowaną wartość

dominanty (różnica pomiędzy oszacowaniami dominanty w obu przypadkach wynosi ponad 600 zł).

Następnie zbadano wrażliwość wyników oszacowanej dominanty metodą standardową dla różnych rozpiętości przedziałów w szeregu rozdzielczym (w zł): 10, 20, 35, 50, 100, 200, 300, 400, 500,37 i 800. Podobnie jak wcześniej dla zadanych rozpiętości przedziałów rozważono dwa warianty początku pierwszego przedziału: równy 0 i równy zaobserwowanej minimalnej wartości cechy, czyli 1763,00 zł (tabl. 3).

Tabl. 3. Wyniki estymacji dominanty na podstawie wzoru interpolacyjnego

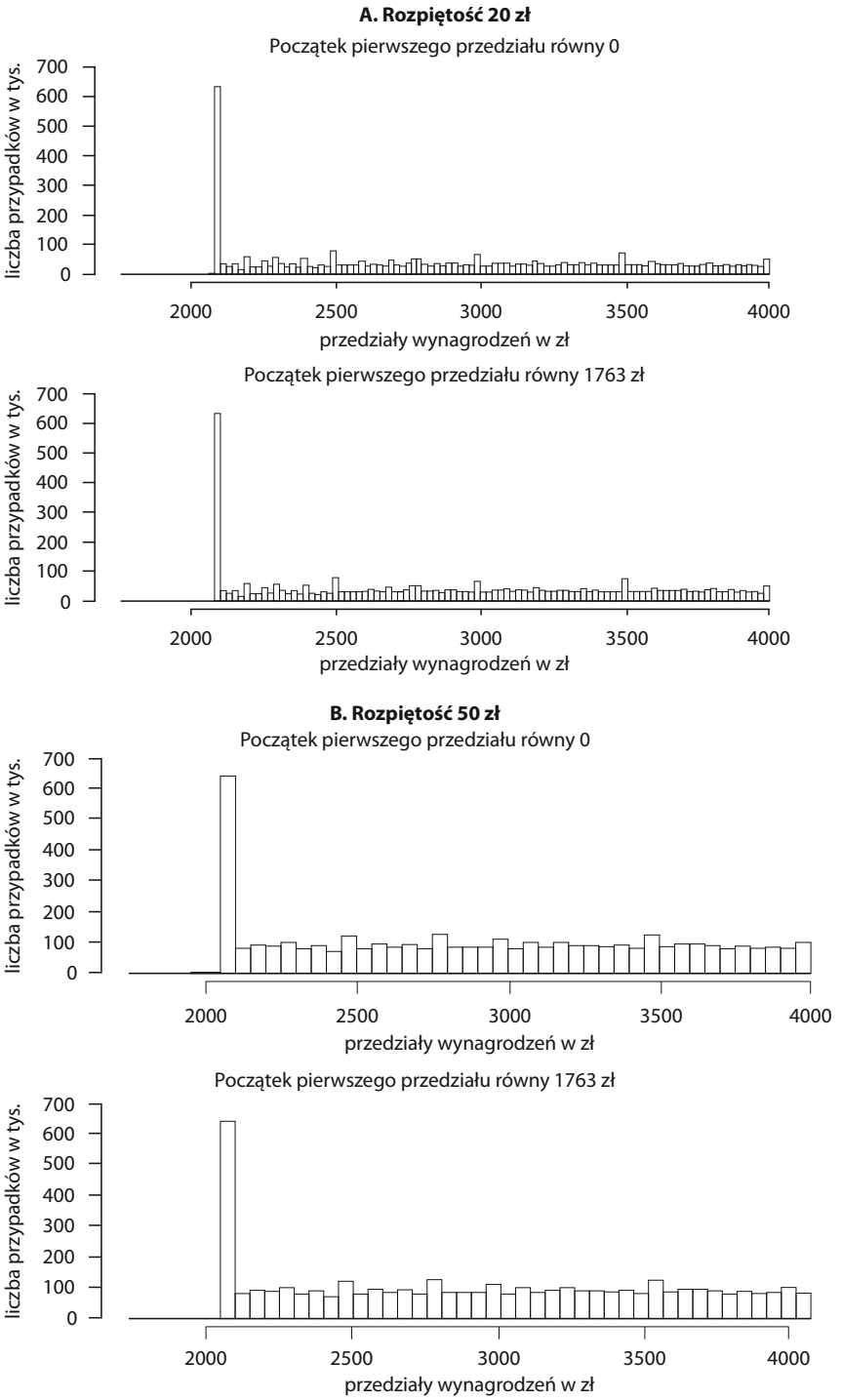
Rozpiętość przedziałów	Przedział dominanty		Oszacowanie dominanty
	od	do	
	w zł		
Początek pierwszego przedziału równy 0			
10	2093,01	2103,00	2095,09
20	2080,01	2100,00	2090,28
35	2065,01	2100,00	2083,24
50	2050,01	2100,00	2076,66
100	2000,01	2100,00	2057,64
200	2000,01	2200,00	2127,92
300	1800,01	2100,00	2054,47
400	2000,01	2400,00	2294,10
500,37	2001,53	2501,90	2379,67
800	2400,01	3200,00	3089,62
Początek pierwszego przedziału równy 1763			
10	2093,01	2103,00	2098,07
20	2083,01	2103,00	2093,23
35	2078,01	2113,00	2096,05
50	2063,01	2113,00	2089,35
100	2063,01	2163,00	2119,72
200	1963,01	2163,00	2094,98
300	2063,01	2363,00	2262,62
400[a]	.		
500,37	2763,77	3264,14	3013,79
800	2563,01	3363,00	2623,21

a Nie można wyznaczyć dominanty (największa liczba przypadków występuje w pierwszym przedziale).
Źródło: obliczenia własne na podstawie: GUS (2020b).

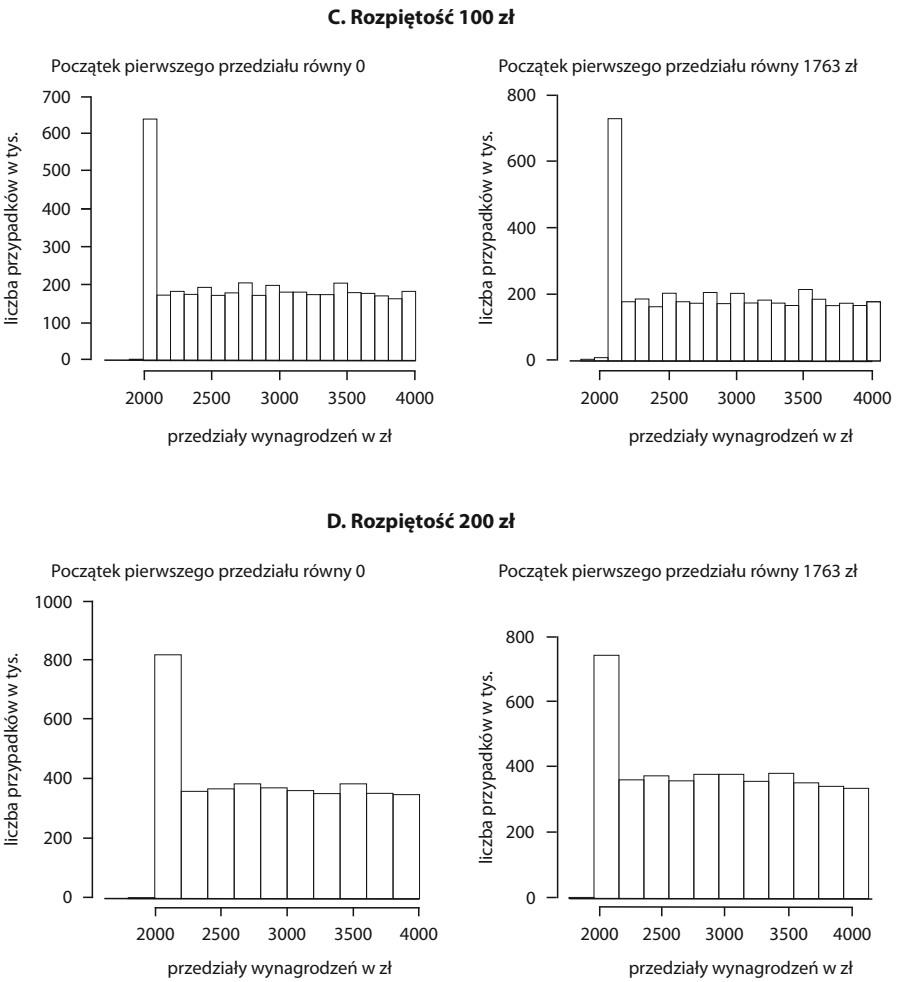
Przybliżony rozkład wynagrodzeń dla wybranych rozpiętości przedziałów uwzględnionych w tabl. 3 przedstawiono za pomocą histogramów (wykr. 2). Z tabl. 3 wynika, że oszacowana dominanta wyznaczona ze wzoru interpolacyjnego oscyluje wokół niskich wynagrodzeń[3], w związku z czym wykr. 2 przedstawia tylko fragmenty histogramów zawierające najliczniejsze przedziały (do wynagrodzenia ok. 4000 zł, a dla rozpiętości 800 zł – do ok. 6000 zł). Histogram przedstawiający całą populację nie byłby czytelny (zob. wykr. 1).

[3] Należy oczekiwać dążenia do wartości ustawowego wynagrodzenia minimalnego.

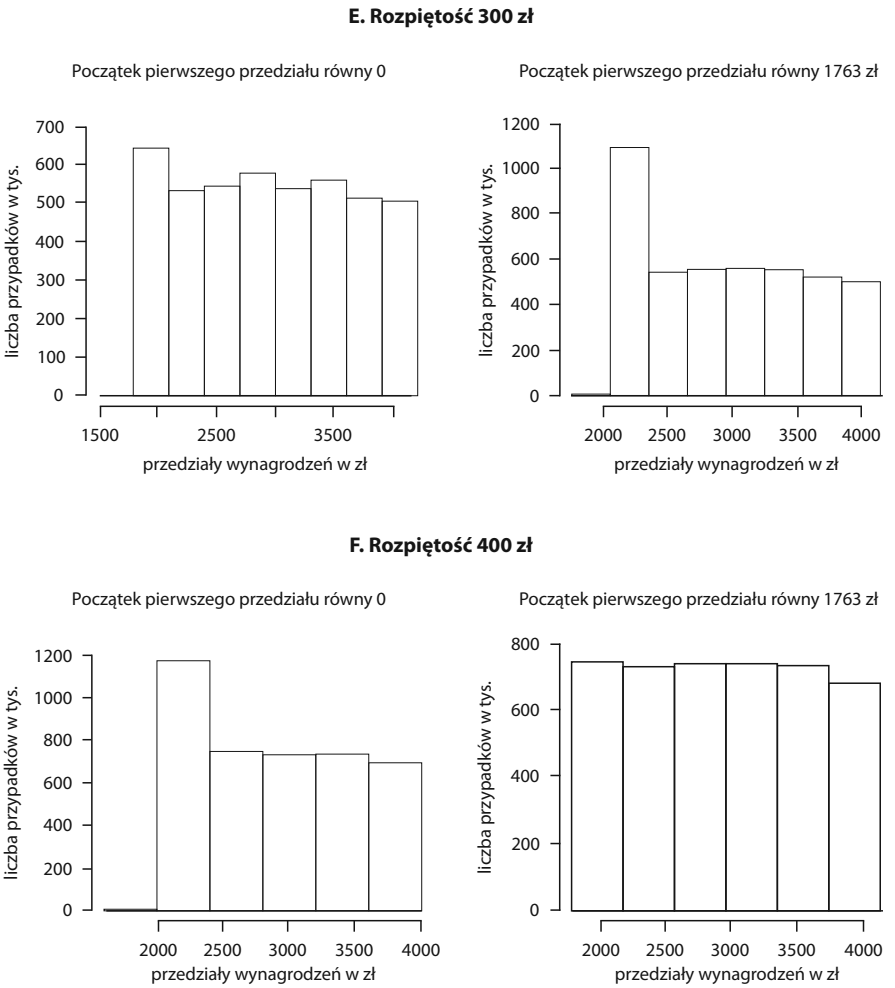
Wykr. 2. Fragmenty histogramu liczby przypadków dla różnych rozpiętości przedziałów



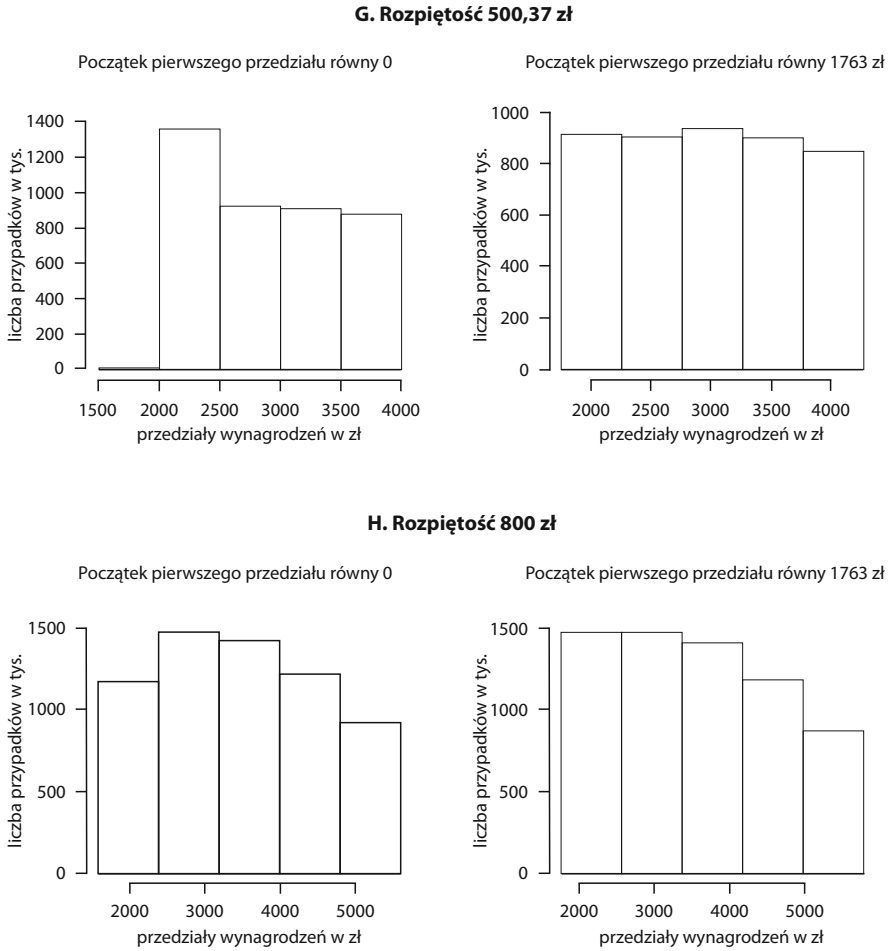
Wykr. 2. Fragmenty histogramu liczby przypadków dla różnych rozpiętości przedziałów (cd.)



Wykr. 2. Fragmenty histogramu liczby przypadków dla różnych rozpiętości przedziałów (cd.)



Wykr. 2. Fragmenty histogramu liczby przypadków dla różnych rozpiętości przedziałów (dok.)



Źródło: obliczenia własne z wykorzystaniem programu R na podstawie: GUS (2020b).

Histogram prezentuje dodatkowe cechy populacji, które mają wpływ na oszacowanie dominanty. Zmniejszenie rozpiętości przedziału w szeregu rozdzielczym powoduje, że wartość oszacowanej dominanty dąży do ustawowej płacy minimalnej, która w 2018 r. wynosiła 2100,00 zł. Przeprowadzono zatem dodatkową analizę liczby przypadków wynagrodzenia równego płacy minimalnej oraz z przedziałów sąsiednich (tabl. 4).

Z analizy wynika, że płacę minimalną otrzymuje 7,44% populacji. Można uznać, że jest to wartość oszacowanej dominanty wynikająca z definicji, przy założeniu, że rozkład wynagrodzeń jest dyskretny. Jednak oszacowana dominanta równa płacy minimalnej reprezentuje bardzo małą frakcję przypadków (osób pobierających wynagrodzenie), więc nie jest w pełni satysfakcjonującym, reprezentatywnym wynikiem. Najlepszym rozwiązaniem byłoby zastosowanie metody pozwalającej na zidentyfikowanie wartości oszacowanej dominanty reprezentującej większą grupę przypadków – o ile takie zjawisko występuje.

W przypadku szeregu rozdzielczego z przedziałami o małej rozpiętości (np. 10, 20, 50 zł) przedziałem oszacowanej dominanty jest zawsze przedział zawierający płacę minimalną.

Tabl. 4. Liczebność w przedziale z płacą minimalną i w przedziałach z nią sąsiadujących

Przedziały	Przedział w zł		Liczba przypadków	Odsetek populacji
	od	do		
Płaca minimalna	2100,00		626 457	7,44
Rozpiętość przedziału 1 zł				
(PM – 1 zł; PM>	2099,01	2100,00	631 917	7,50
(PM; PM + 1 zł>	2100,01	2101,00	5088	0,06
Rozpiętość przedziału 10 zł				
(PM – 10 zł; PM>	2090,01	2100,00	632 339	7,51
(PM; PM + 10 zł>	2100,01	2110,00	21 454	0,25
Rozpiętość przedziału 11 zł				
(PM – 11 zł; PM>	2089,01	2100,00	632 387	7,51
(PM; PM + 11 zł>	2100,01	2111,00	22 347	0,27

Uwaga. PM – płaca minimalna.

Źródło: obliczenia własne na podstawie: GUS (2020b).

Wnioski z zastosowania metody standardowej, wykorzystującej wzór interpolacyjny, są następujące:

- dominanta wyznaczona dla badanej cechy jest wrażliwa na przyjętą rozpiętość przedziałów w szeregu rozdzielczym oraz na początek pierwszego przedziału (różnica pomiędzy skrajnymi wynikami obliczeń mody wynosi ok. 1000 zł);
- zmniejszanie rozpiętości przedziałów powoduje, że oszacowana dominanta dąży do ustawowej płacy minimalnej (2100,00 zł), która jest wartością „definicyjną" przybliżonej dominanty, ale reprezentuje małą frakcję przypadków;

- na wykr. 2 dla przedziałów zawierających okrągłe wartości wynagrodzeń będące wielokrotnością 100 zł lub 500 zł można zauważyć pewną powtarzalność (cykl), np. co ok. 100 zł czy 500 zł, co oznacza, że zawiera się więcej umów na takie okrągłe kwoty wynagrodzeń niż umów na inne kwoty;
- dolna granica pierwszego przedziału równa minimum badanej cechy wydaje się właściwa, biorąc pod uwagę warunki, jakie należy spełnić, aby zastosować wzór interpolacyjny (1), a także uwzględniając zasady grupowania obserwacji w szereg rozdzielczy;
- w przypadku większych rozpiętości przedziałów oszacowany rozkład w sąsiedztwie dominanty jest bardziej spłaszczony (wykr. 2G i 2H), ale precyzja oszacowań maleje. Przedziały sąsiadujące z przedziałem dominanty mają zbliżoną liczebność, więc wiarygodność wyników jest wątpliwa;
- metoda standardowa nie pozwala na znalezienie takiej wartości dominanty, która reprezentuje znaczną frakcję przypadków.

3.2. Estymacja dominanty za pomocą zaawansowanych metod statystycznych

Na drugim etapie badania zastosowano zaawansowane nieparametryczne metody estymacji dominanty, dostępne w dodatku modeest w pakiecie R[4] (tabl. 5). Zgodnie z klasyfikacją zastosowaną w części 2.2 można wśród nich wyróżnić takie, które wykorzystują numeryczną optymalizację funkcji gęstości i estymację jądrową. Są to metody: Parzena, Tsybakova i Meanshift. Pozostałe są metodami iteracyjnymi, które nie wykorzystują estymacji jądrowej.

Tabl. 5. Zaawansowane metody estymacji dominanty wynagrodzeń i jej oszacowana wartość

Metoda	Oszacowana wartość dominanty w zł
Naiwna	3560,12
Lientza: beta = 0,2[a]	2948,62
beta = 0,3[a]	3002,46
Ventera (bw = 1/3)[b]	2571,74
Ventera	2946,46
Grenandera	3805,38
HSM	2100,00[c]
Parzena (kernel = „uniform")[d]	3560,12
Asselina	2745,83
Tsybakova (kernel = „gaussian")[d]	2946,46
Meanshift	3015,52

a Parametr beta w funkcji *F* (Lientz, 1970). b bw (bandwith) – parametr opisujący szerokość pasma w metodzie Ventera (Poncet, 2019). c „The distribution could be multimodal" (rozkład może być wielomodalny) – uwaga wygenerowana przez program R. d Kernel oznacza rodzaj jądra użytego w danej metodzie (Poncet, 2019).

Źródło: obliczenia własne z wykorzystaniem dodatku modeest w programie R na podstawie: GUS (2020b).

[4] Szczegółowy opis metod oraz przywołania literatury znajdują się w opracowaniu Ponceta (2019).

Wnioski z zastosowania zaawansowanych metod estymacji dominanty są następujące:

- dominanta z próby oszacowana za pomocą metod przedstawionych w tabl. 5 przyjmuje wartości od ok. 2100 zł do ok. 3800 zł;
- metoda HSM szacuje dominantę na poziomie płacy minimalnej, a więc wskazuje wartość, która wynika z definicji modalnej jako wartości najczęstszej;
- za pomocą metody Grenandera uzyskuje się najwyższy szacunek dominanty;
- większość zastosowanych metod szacuje dominantę na poziomie od ok. 2700 zł do ok. 3500 zł, a zatem rozrzut jest znaczny (800 zł);
- oszacowania dominanty otrzymane za pomocą analizowanych zaawansowanych metod statystycznych nie dążą do jednej, zbliżonej wartości dominanty wynagrodzeń, która ma uzasadnienie społeczno-gospodarcze;
- ze względu na duży rozrzut otrzymanych wyników można przypuszczać, że metody wymienione w tabl. 5 nie są odpowiednie w przypadku rozkładów skrajnie asymetrycznych, charakteryzujących się dodatkowo pewną cyklicznością (których przykładem jest rozkład wynagrodzeń w Polsce).

4. Podsumowanie

Dominanta wynagrodzeń jest ważną miarą opisującą ich strukturę, ale ze względu na silną asymetrię rozkładu (skupienie blisko płacy minimalnej) jej estymacji nie wykonuje się standardowo w analizie wynagrodzeń. W niniejszym artykule dokonano przeglądu metod estymacji dominanty i podjęto próbę oszacowania dominanty wynagrodzeń brutto w Polsce, których wartości zaczerpnięto z reprezentacyjnego badania struktury wynagrodzeń według zawodów za październik 2018 r. przeprowadzonego przez GUS w przedsiębiorstwach o liczbie pracujących powyżej dziewięciu osób (GUS, 2020b). Najpierw zastosowano najprostszą metodę (obliczanie ze wzoru interpolacyjnego) przy różnych założeniach dotyczących rozpiętości przedziałów oraz początku pierwszego przedziału. Następnie oszacowano dominantę wynagrodzeń za pomocą zaawansowanych metod statystycznych.

Na podstawie otrzymanych wyników można zauważyć, że różne metody warunkują różne oszacowania dominanty. W przypadku metody standardowej jej wartość jest wrażliwa na założoną rozpiętość przedziałów w szeregu rozdzielczym i początek pierwszego przedziału. Zmniejszanie rozpiętości przedziałów powoduje, że wartość dominanty dąży do wartości równej płacy minimalnej, a zatem wskazuje przybliżenie dominanty wynikające bezpośrednio z definicji, ale reprezentujące bardzo małą frakcję przypadków (osób pobierających wynagrodzenie). Zaawansowane metody statystyczne również mają ograniczone zastosowanie do analizowanych danych, ponieważ wyniki różnią się znacznie w zależności od zastosowanej metody.

Analiza otrzymanych wyników daje podstawy do rozważenia tezy, że rozkład wynagrodzeń jest mieszany: ma charakter rozkładu dyskretnego dla wynagrodzeń w wysokości płacy minimalnej i rozkładu ciągłego dla wynagrodzeń powyżej płacy minimalnej. Dodatkowo można zauważyć pewną powtarzalność (cykl), np. co ok. 100 zł czy 500 zł, co oznacza, że w Polsce zawiera się więcej umów, w których kwota wynagrodzenia jest okrągła, niż umów na inne kwoty.

Posługiwanie się dominantą do opisywania struktury wynagrodzeń nie ma zatem zastosowania w przypadku zbioru o bardzo dużej liczbie obserwacji (danych quasi-ciągłych), którego rozkład jest skrajnie asymetryczny oraz w którym występują zaburzenia (wynikające np. z popularności okrągłych wartości wynagrodzeń lub innych fluktuacji). Zarówno standardowa metoda interpolacji, jak i bardziej zaawansowane numerycznie metody estymacji rozkładu nie pozwalają wyznaczyć wartości dominanty, która opisuje rozkład wynagrodzeń w oczekiwany sposób, tzn. tak, aby reprezentowała odpowiednio wysoką frakcję przypadków oraz dawała zbliżone wyniki (arbitralnie np. ±50 zł czy ±100 zł) dla różnych metod i założeń.

Wydaje się, że rozkład wynagrodzeń w Polsce jest dwumodalny[5]. Jedna moda występuje w pobliżu wynagrodzenia minimalnego i jest wskazywana zarówno przy bezpośrednim zastosowaniu definicji, jak i przy wykorzystaniu histogramów o wąskich przedziałach. Druga moda (dużo bardziej wypłaszczona) jest widoczna na histogramach dla przedziałów o większych rozpiętościach i wyznaczana numerycznie za pomocą większości zaawansowanych metod statystycznych. Analiza pomija szczegóły drobniejszych wahań cyklicznych spowodowanych popularnością okrągłych kwot wynagrodzeń.

W przypadku zmiennych o rozkładach mieszanych wyznaczanie dominanty jest skomplikowane i wymaga niestandardowego podejścia. Będzie to przedmiotem dalszych prac badawczych autorów.

Bibliografia

Bickel, D. R. (2002). Robust estimators of the mode and skewness of continuous data. *Computational Statistics and Data Analysis*, *39*(2), 153–163. https://doi.org/10.1016/S0167-9473(01)00057-3.

Bickel, D. R. (2003). Robust and efficient estimation of the mode of continuous data: The mode as a viable measure of central tendency. *Journal of Statistical Computation and Simulation*, *73*(12), 899–912. https://doi.org/10.1080/0094965031000097809.

Bickel, D. R, Frühwirth, R. (2006). On a Fast, Robust Estimator of the Mode: Comparisons to Other Robust Estimators with Applications. *Computational Statistics and Data Analysis*, *50*(12), 3500–3530. https://doi.org/10.1016/j.csda.2005.07.011.

[5] W danych można jeszcze spodziewać się pewnego rodzaju anomalii związanych z tendencją do wyższej liczebności w przypadku okrągłych wartości wynagrodzeń (wielokrotności 100 zł czy 500 zł). Można jednak założyć, że nie zaburza to ogólnego obrazu i wniosków.

Fukunaga, K., Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, *21*(1), 32–40. https://doi.org/10.1109/TIT.1975.1055330.

Grenander, U. (1965). Some direct estimates of the mode. *Annals of Mathematical Statistics*, *36*(1), 131–138.

Główny Urząd Statystyczny. (2020a). *Podstawowe miary statystyczne. Miary położenia*. https://eks.stat.gov.pl/materialy/scenariusze/miary_statystyczne/materialy_dla_nauczyciela.pdf.

Główny Urząd Statystyczny. (2020b). *Struktura wynagrodzeń według zawodów w październiku 2018 r.* https://stat.gov.pl/obszary-tematyczne/rynek-pracy/pracujacy-zatrudnieni-wynagrodzenia-koszty-pracy/struktura-wynagrodzen-wedlug-zawodow-w-pazdzierniku-2018-roku,4,9.html.

Główny Urząd Statystyczny. (2020c). *Zeszyt metodologiczny. Struktura wynagrodzeń według zawodów*. https://stat.gov.pl/obszary-tematyczne/rynek-pracy/zasady-metodyczne-rocznik-pracy/zeszyt-metodologiczny-struktura-wynagrodzen-wedlug-zawodow,8,1.html.

Główny Urząd Statystyczny. (2021). *Struktura wynagrodzeń według zawodów w październiku 2020 r.* https://stat.gov.pl/obszary-tematyczne/rynek-pracy/pracujacy-zatrudnieni-wynagrodzenia-koszty-pracy/struktura-wynagrodzen-wedlug-zawodow-w-pazdzierniku-2020-roku,5,7.html.

Jurkiewicz, T., Kozłowski, A. (2009). O wyznaczaniu dominanty rozkładu cechy ciągłej w szeregach szczegółowych. *Acta Universitatis Lodziensis. Folia Oeconomica*, *227*, 107–120.

Lientz, B. P. (1970). Results on nonparametric modal intervals. *SIAM Journal of Applied Mathematics*, *19*(2), 356– 366. https://doi.org/10.1137/0119034.

Lientz, B. P. (1972). Properties of modal intervals. *SIAM Journal of Applied Mathematics*, *23*(1), 1–5. https://doi.org/10.1137/0123001.

Ostasiewicz, S., Rusnak, Z., Siedlecka, U. (1995). *Statystyka. Elementy teorii i zadania*. Wydawnictwo Akademii Ekonomicznej we Wrocławiu.

Parlińska, M., Parliński, J. (2011). *Statystyczna analiza danych z Excelem*. Wydawnictwo SGGW.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076.

Poncet, P. (2019, 18 stycznia). *Mode Estimation.* https://github.com/paulponcet/modeest.

Sokołowski, A. (2013). *Bezpośrednie estymatory modalnej.* Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Stanisławek, J. (2010). *Podstawy statystyki. Opis statystyczny. Korelacja i regresja. Rozkłady zmiennej losowej. Wnioskowanie statystyczne*. Oficyna Wydawnicza Politechniki Warszawskiej.

Tsybakov, A. B. (1990). Recursive estimation of the mode of a multivariate distribution. *Problems of Information Transmission*, *26*(1), 31–37.

Venter, J. H. (1967). On estimation of the mode. *The Annals of Mathematical Statistics*, *38*(5), 1446–1455.