Małgorzata Kobylińska[1]

Kontury zanurzania i krzywa skali w ocenie zróżnicowania handlu elektronicznego w krajach Unii Europejskiej

1. Wstęp

Koncepcja zanurzania obserwacji w próbie zapoczątkowana została przez J.W. Tukeya (1975), który zdefiniował pojęcie półprzestrzeni zanurzania. Zagadnienia dotyczące zanurzania obserwacji w próbie, poruszane coraz częściej w literaturze przedmiotu, odnoszą się zarówno do kwestii teoretycznych, jak i aplikacyjnych. W pracy D.L. Donoha i M. Gasko (1992) zdefiniowano pojęcia dotyczące zanurzania obserwacji w próbie oraz przedstawiono własności półprzestrzeni zanurzania Tukeya. Wykorzystanie koncepcji zanurzania w analizie danych wielowymiarowych zaprezentowano między innymi w pracach, których autorami są: R.Y. Liu i in. (2006), Mosler (2002, 2013) lub Kobylińska (2017).

Celem pracy jest zaprezentowanie wykorzystania konturów zanurzania obserwacji w próbie oraz krzywej skali w badaniu zróżnicowania danych wielowymiarowych. Posłużono się danymi liczbowymi dotyczącymi handlu elektronicznego w krajach UE. Zastosowanie odpowiedniego oprogramowania komputerowego umożliwiło dokonanie obliczeń oraz wykonanie wykresów wykorzystanych w artykule.

W literaturze spotkać można wiele definicji handlu elektronicznego. Według J. Wielkiego (2000) jest on rozumiany jako sprzedaż oraz zakup towarów, usług i informacji wyłącznie za pomocą sieci komputerowych.

W ostatnich latach zauważyć można wzrost obrotów odnoszących się do handlu elektronicznego. Jest on bardziej popularny wśród młodszego pokolenia, które stanowi dwie trzecie osób dokonujących zakupów online[2]. Spowodowane

[1] Uniwersytet Warmińsko-Mazurski w Olsztynie, Wydział Nauk Ekonomicznych.

[2] https://www.ecommerce-europe.eu/press-item/european-ecommerce-report-2017-released-ecommerce-continues-prosper-europe-markets-grow-different-speeds/#_ftnref1 (dostęp: 10.07.2018).

jest to rozwojem technologii informatycznych, coraz szerszym dostępem do technologii bezprzewodowej oraz urządzeń mobilnych.

2. Metody badawcze

Zanurzanie obserwacji w próbie może być wykorzystywane do porządkowania obserwacji wielowymiarowych względem oddalenia od centrum zbioru danych, które jest wyznaczone przez medianę Tukeya, czyli punkt, któremu odpowiada najwyższa wartość zanurzania. Przy wykorzystaniu porządkowania obserwacji względem centralnego skupiania możliwe jest określenie pewnych własności analizowanych zbiorów danych m.in.: położenia, rozproszenia lub skośności. Ważną zaletą tych metod jest odporność na występowanie obserwacji odstających, które mogą pojawić się podczas badania zjawisk ekonomicznych[3].

Metody wykorzystane w pracy wymagają zdefiniowania pewnych pojęć związanych z koncepcją zanurzania obserwacji w próbie. W artykule analiza przeprowadzona została dla przypadku dwuwymiarowego, w związku z tym poniższe definicje zostaną przedstawione tylko dla tego wymiaru.

Załóżmy, że dana jest próba dwuwymiarowa P_n^2 o liczebności n, oraz że punkt $\theta \in R^2$. Każdy punkt próby x_{ij} pojmowany jest jako obserwacja j-tej zmiennej zaobserwowana na i-tym obiekcie ($i = 1, 2, \ldots, n$, $j = 1,2$). Każdą obserwację próby dwuwymiarowej P_n^2 zapisać można w postaci wektora $x_i = \left[x_{i1}, x_{i2}\right]$.

Niech k określa minimalną liczbę punktów próby P_n^2 należącą do zamkniętej półpłaszczyzny, której linia rozdzielająca przechodzi przez punkt θ. Punkt θ może być elementem próby P_n^2 lub dowolnym punktem z przestrzeni rzeczywistej R^2.

Półpłaszczyznę zanurzania Tukeya punktu $\theta \in R^2$ względem rozkładu prawdopodobieństwa P_n^2 na R^2 definiujemy jako minimalne prawdopodobieństwo, skoncentrowane na domkniętej półpłaszczyźnie, której linia brzegowa przechodzi przez punkt θ. Kontury zanurzania próby dwuwymiarowej P_n^2 wyznaczone są na podstawie półpłaszczyzny zanurzania Tukeya[4].

Własności półprzestrzeni zanurzania opisane zostały między innymi w pracy Y. Zua i R. Serflinga (2000).

---

[3] D.L. Donoho, M. Gasko, *Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness*, „The Annals of Statistics" 1992, 20, s. 1803–1827.

[4] Y. Zuo, R. Serfling, *General Notions of Statistical Depth Function*, „The Annals of Statistics" 2000, 28, s. 461–482, http://dx.doi.org/10.1214/aos/1016218226 (dostęp: 5.07.2018).

Konturem zanurzania stopnia k-tego nazywamy zbiór:

$$Con_k = \left\{\theta : zan_2\left(\theta, P_n^2\right) = k\right\} \tag{1}$$

dla $k = 1, 2, \ldots[n/2]$, gdzie $[n/2]$ oznacza część całkowitą liczby $n/2$, $zan_2(\theta, P_n^2)$ – zanurzanie punktu θ w próbie P_n^2.

Przynależność obserwacji do konturów zanurzania umożliwia dokonanie ich rangowania względem oddalenia od centrum próby. Obserwacja należąca do konturu o najwyższym stopniu zanurzania określana jest mianem dwuwymiarowej mediany Tukeya. Jeżeli kilka obserwacji spełnia ten warunek, mediana wyznaczona jest jako ich środek ciężkości[5].

Obszarem ograniczonym przez kontur o k-tym stopniu zanurzania nazywamy zbiór:

$$OCon_k\left\{\theta; zan_2\left(\theta, P_n^p\right) \geq k\right\} \tag{2}$$

Zbiór określony jako:

$$Con_p = \bigcap_k \left\{OCon_k : P(OCon_k) \geq p\right\} \tag{3}$$

nazywamy p-tym obszarem centralnym, czyli najmniejszym obszarem domkniętym i ograniczonym przez kontur o k-tym stopniu zanurzania, zawierającym $\lceil np \rceil$ najbardziej centralnych punktów próby, $p \in \langle 0,1 \rangle$, n jest liczebnością próby P_n^2, symbol $\lceil A \rceil$ określa część całkowitą liczby A[6].

Kontury zanurzania wykorzystywane w pracy obliczone zostały zgodnie z algorytmem „isodepth", który szczegółowo opisano w pracy I. Ruts i P.J. Rousseeuwa (1996). Wykresy konturów zanurzania wykonano, wykorzystując pakiet „depth" środowiska R, autorstwa M. Genesta, J.C. Masse'a i J.F. Plante'go[7]. Pakiet „DepthProc" autorstwa D. Kosiorowskiego, M. Bociana, A. Wegrzynowskiej i Z. Zawadzkiego wykorzystano do sporządzenia wykresu krzywej skali[8].

Kontury zanurzania obserwacji w próbie pozwalają na uchwycenie pewnych własności zbiorów danych, m.in. położenia rozkładu (wektor medianowy),

---

[5] I. Ruts, P.J. Rousseeuw, *Isodepth: A Program for Depth Contours*, Proceedings in Computational Statistics, A. Prat (red.), COMPSTAT 1996, Physica, Heidelberg, s. 441–446.

[6] R.Y. Liu, J.M. Parelius, K. Singh, *Multivariate Analysis by Data Depth: Descriptive statistics, Graphics and Inference (with Discussion)*, „The Annals of Statistics" 1999, 27, s. 783–858.

[7] https://cran.r-project.org/web/packages/depth/depth.pdf (dostęp: 15.06.2018).

[8] https://cran.r-project.org/web/packages/DepthProc/index.html (dostęp: 15.06.2018).

korelacji pomiędzy zmiennymi (orientacja wykresów konturów zanurzania), asymetrii oraz koncentracji rozważanych zmiennych.

W pracy R.Y. Liu i in. (1999) zaproponowano narzędzie umożliwiające porównanie zróżnicowania rozkładów wielowymiarowych. Oparte jest ono na wykresie krzywej skali, gdzie na osi X umieszczono wartości prawdopodobieństwa $p \in \langle 0,1 \rangle$, a na osi Y odpowiadające im wartości objętości p-tych obszarów centralnych wyznaczonych dla danych zbiorów.

Szybkość wzrostu objętości p-tych obszarów centralnych w stosunku do wartości prawdopodobieństwa umożliwia ocenę zróżnicowania zbioru danych. Jeżeli dla danej wartości prawdopodobieństwa objętość p-tego obszaru centralnego rozkładu F jest większa od objętości p-tego obszaru centralnego rozkładu G (wykres krzywej skali rozkładu F znajduje się powyżej wykresu dla rozkładu G) można przyjąć, że rozkład F charakteryzuje się większym zróżnicowaniem wartości badanych zmiennych względem centralnego skupienia w porównaniu z rozkładem G. Szybciej wzrastające wartości objętości p-tych obszarów centralnych w porównaniu z wartościami p świadczą o większym rozproszeniu punktów w badanym zbiorze danych[9].

3. Wyniki badań

Dane wykorzystane w pracy zostały zaczerpnięte z rejestru Eurostatu z lat 2014 i 2016[10]. Analizie poddano kraje Unii Europejskiej ze względu na następujące zmienne:

- X_1 – odsetek osób, które korzystały z Internetu w ciągu ostatnich trzech miesięcy w celu sprzedaży towarów i usług (w %),
- X_2 – odsetek osób, które korzystały z Internetu w ciągu ostatnich trzech miesięcy i w tym czasie dokonały zakupu online (w %).

Wartości badanych zmiennych zamieszczone zostały w tabeli 1.

Zauważyć można różnicę pomiędzy odsetkami osób korzystających z Internetu w celu sprzedaży towarów i usług oraz dokonujących zakupów online. W badanych latach w krajach UE średni odsetek osób korzystających z Internetu w celu sprzedaży towarów i usług nie przekroczył 20% i wynosił odpowiednio 19,14 oraz 18,21%. Większą popularnością cieszyło się dokonywanie zakupów

[9] R.Y. Liu, J.M. Parelius, K. Singh, op. cit., s. 783–858

[10] http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do (dostęp: 10.06.2018).

online. Przeciętny odsetek osób w UE był w tym przypadku ponad dwukrotnie wyższy i sięgał odpowiednio 43,75 oraz 47,04%.

Tabela 1. Odsetek osób, które korzystały z Internetu w krajach UE w celu sprzedaży towarów i usług oraz dokonały zakupu online w 2014 i 2016 r.

Lp.	Kraj	2014		2016	
		X_1	X_2	X_1	X_2
1	Austria	14	54	13	57
2	Belgia	23	48	24	54
3	Bułgaria	12	18	11	19
4	Chorwacja	31	31	38	35
5	Cypr	1	33	6	29
6	Czechy	19	32	15	35
7	Dania	27	69	36	73
8	Estonia	29	44	22	52
9	Finlandia	24	57	22	51
10	Francja	35	58	28	61
11	Grecja	6	31	3	34
12	Hiszpania	13	36	15	43
13	Holandia	31	63	36	67
14	Irlandia	14	54	13	50
15	Litwa	5	26	7	33
16	Luksemburg	15	66	15	70
17	Łotwa	5	31	7	39
18	Malta	23	56	27	53
19	Niemcy	32	71	33	72
20	Polska	17	37	21	42
21	Portugalia	11	27	11	33
22	Rumunia	3	11	5	13
23	Słowacja	14	39	16	51
24	Słowenia	40	36	22	39
25	Szwecja	16	66	19	67
26	Węgry	24	27	14	34
27	Wielka Brytania	37	79	22	82
28	Włochy	15	25	9	29

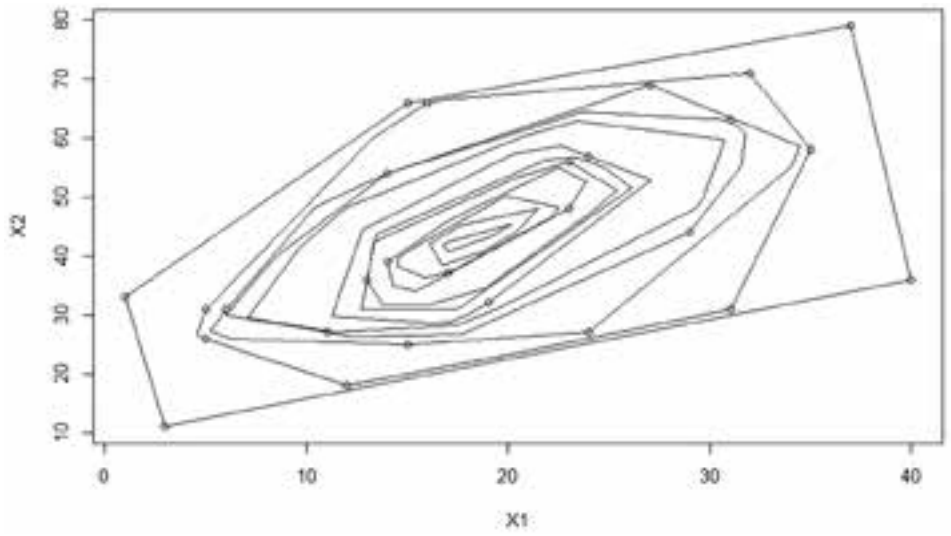
Źródło: Eurostat.

W danych latach Cypr, Grecja, Litwa, Łotwa oraz Rumunia charakteryzowały się najniższym odsetkiem osób, które korzystały z Internetu w celu sprzedaży towarów i usług. Nie przekroczył on w tym przypadku 10%. W 2016 r. do krajów, w których odsetek ten wynosił ponad 30%, należały Chorwacja, Dania, Holandia oraz Niemcy. W porównaniu z 2014 r. w UE odsetek ten zmniejszył się przeciętnie o 0,93 p.p.

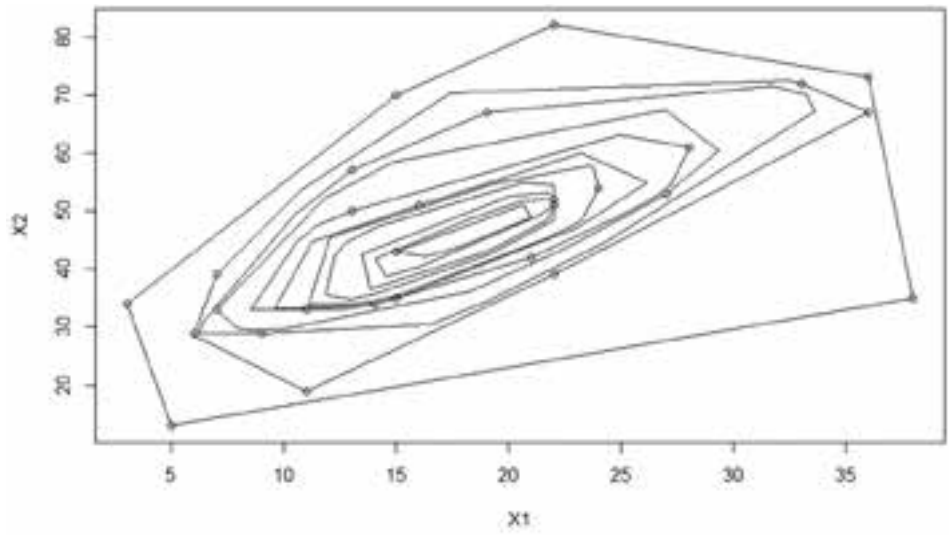
Tylko w Wielkiej Brytanii w 2016 r. odsetek osób korzystających z Internetu, które dokonały zakupów online, wyniósł ponad 80% (82%). W porównaniu z Rumunią, w której zanotowano najniższą wartość tej cechy, był o 69 p.p. wyższy. Najmniej osób dokonywało zakupów online w Bułgarii oraz w Rumunii, odsetek w tym przypadku nie przekroczył 20%. W porównaniu z 2014 r. w krajach UE wzrósł on o 3,29 p.p. Wartości badanych zmiennych uplasowały Polskę odpowiednio na 14. i 13. miejscach w przypadku zakupów online oraz na 15. i 17. miejscach w przypadku drugiej zmiennej.

Zakres zmienności dla zakupów online wynosił odpowiednio 68 i 69% i był prawie dwukrotnie wyższy w stosunku do odsetka osób, które dokonały sprzedaży towarów i usług z wykorzystaniem Internetu (odpowiednio 38 i 35%). Wartość współczynnika zmienności w przypadku zakupów online w 2016 r. wynosiła 36,77%, co wskazuje na umiarkowane zróżnicowanie cechy. W przypadku pozostałych zmiennych jego wartość przekroczyła 40% i można stwierdzić, że zróżnicowanie wartości tych zmiennych było silne.

Wykresy konturów zanurzania dla analizowanych zbiorów danych przedstawiono na rysunkach 1 i 2. Obszary ograniczone konturami o danym stopniu zanurzania są wypukłe i wstępujące, tj. spełniają warunek $OCon_k \subseteq OCon_{k-1}$. Wektory medianowe zostały wyznaczone jako środki ciężkości wierzchołków konturów z najwyższą wartością zanurzania Tukeya. Dla danych z lat 2014 i 2016 skonstruowanych zostało odpowiednio 12 i 11 konturów zanurzania. Dwuwymiarowe wektory medianowe wynoszą $Me_{2014} = \left[18{,}15;42{,}79\right]$ oraz $Me_{2016} = \left[18{,}23;46{,}24\right]$, gdzie wartości miary zanurzania Tukeya dla tych wektorów wynoszą odpowiednio 0,43 i 0,39. Wyższa wartość miary zanurzania Tukeya w przypadku danych z 2014 r. wynika z większej liczby skonstruowanych konturów zanurzania dla tych danych. Miara zanurzania Tukeya wyznaczana jest jako iloraz stopnia konturu, do którego należy dana obserwacja i liczby wszystkich konturów zanurzania, jakie zostały skonstruowane dla danego zbioru danych. Zauważyć można, że żadna obserwacja analizowanych zbiorów nie należy do konturów o najwyższym stopniu zanurzania.



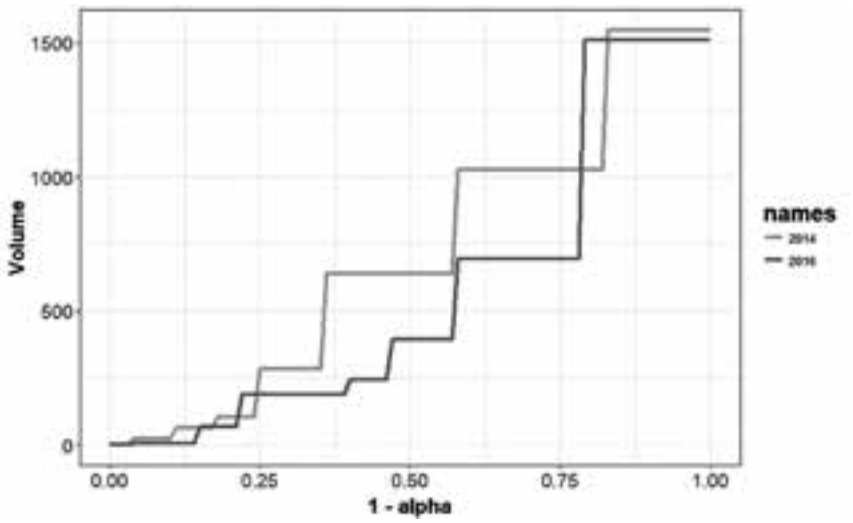
Rysunek 1. Wykres konturów zanurzania dla danych z 2014 r.
Źródło: opracowanie własne z wykorzystaniem pakietu „depth".



Rysunek 2. Wykres konturów zanurzania dla danych z 2016 r.
Źródło: opracowanie własne z wykorzystaniem pakietu „depth".

Zauważyć można, że kształt konturów zanurzania dla 2014 r. jest „bardziej okrągły". W tym roku zaobserwować można mniejszą koncentrację wartości badanych zmiennych wokół wektora medianowego w porównaniu z 2016 r.

Wszystkie wierzchołki powłoki wypukłej (konturu o najniższym stopniu zanurzania Tukeya) wyznaczone zostały przez obserwacje należące do analizowanych zbiorów danych. Wierzchołkami powłok wypukłych jest odpowiednio 5 oraz 6 obserwacji, co stanowi w danych latach 17,86 oraz 21,43% obserwacji. Do wierzchołków powłok wypukłych w każdym z badanych lat należą Luksemburg oraz Wielka Brytania, ze względu na wysoki odsetek osób, które dokonały sprzedaży oraz zakupów z wykorzystaniem Internetu. W Rumunii, należącej do powłoki wypukłej w każdym z badanych lat, odsetek osób, które korzystały z Internetu w celu sprzedaży towarów i usług, nie przekroczył 5%. Spośród wszystkich krajów UE tylko w Bułgarii i Rumunii mniej niż co piąta osoba dokonała zakupu online (w Bułgarii 18 i 19%, natomiast w Rumunii 11% i 13%).



Rysunek 3. Wykres krzywej skali dla danych z lat 2014 i 2016
Źródło: opracowanie własne z wykorzystaniem pakietu „DepthProc".

Wykres krzywej skali (rysunek 3) pozwala na zobrazowanie różnic w rozproszeniu obserwacji zbiorów danych. Dla wartości prawdopodobieństwa od 0,25 do 0,75 wykres dla 2014 r. położony jest wyżej. Wartości pól *p*-tych obszarów centralnych, zawierających 25% lub więcej centralnie położonych obserwacji w przypadku tych danych, są większe. Świadczy to o mniejszej koncentracji obserwacji w tym roku wokół centralnego skupienia w porównaniu z 2016 r. Wartości na osi Y, dla prawdopodobieństwa równego jeden, odpowiadają polom powierzchni powłok wypukłych zbiorów danych. Pola obliczone zostały dla figur wypukłych, zawierających wszystkie obserwacje zbiorów. Zauważyć można,

że różnica pól powłok wypukłych jest mniejsza w porównaniu z różnicą wartości pól *p*-tych obszarów centralnych dla mniejszych wartości prawdopodobieństwa. Jest to wynikiem większej koncentracji obserwacji z 2016 r. wokół wektora medianowego.

4. Podsumowanie

Koncepcja zanurzania obserwacji w próbie może stać się użytecznym narzędziem w analizie danych wielowymiarowych. Kontury zanurzania mogą być wykorzystywane w celu zobrazowania kształtu i orientacji zbiorów danych. Wykresy krzywych skali wykonane na płaszczyźnie umożliwiają łatwą interpretację oraz porównanie rozproszenia analizowanych danych.

Na podstawie przeprowadzonych analiz zauważyć można, że handel w krajach Unii Europejskiej zdominowany został przez takie kraje, jak Dania, Holandia, Luksemburg, Francja, Niemcy, Szwecja oraz Wielka Brytania. Odsetek osób, które dokonały zakupów online, przekroczył 60%. Co najmniej co piąta osoba korzystała z Internetu w celu sprzedaży towarów i usług. Rozwój handlu elektronicznego różni się w zależności od regionu Unii Europejskiej. W krajach zachodniej oraz północno-zachodniej UE odnotowano najwyższy odsetek osób korzystających z handlu elektronicznego, w części południowej i wschodniej wartości tego odsetka były najniższe.

Przedstawione w artykule metody analizy danych oparte na koncepcji zanurzania obserwacji w próbie mogą być wykorzystywane w analizie danych wielowymiarowych, mogą stać się uzupełnieniem lub alternatywą dla metod klasycznych, wykorzystywanych w celu analizy dyspersji lub korelacji analizowanych zmiennych.

Bibliografia

Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications, R. Liu, R. Serfling, D.L. Souvaine (red.), DIMACS Series 2006, vol. 72, American Mathematical Society.

Donoho D.L., Gasko M., *Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness*, „The Annals of Statistics" 1992, 20, s. 1803–1827.

Kobylińska M., *Zanurzanie obserwacji w próbie w ocenie zróżnicowania przestępczości przeciwko mieniu oraz stopy bezrobocia w Polsce*, „Metody Ilościowe w Badaniach Ekonomicznych" 2017, vol. XVIII, s. 602–613.

Liu R.Y., Parelius J.M., Singh K., *Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference*, „The Annals of Statistics" 1999, 27, s. 783–858.

Mosler K., *Multivariate Dispersion, Central Regions and Depth: The Lift Zonoid Approach*, Springer, New York 2002.

Mosler K., *Depth statistics*, 2013, http://arxiv.org/pdf/1207.4988.pdf

Ruts I., Rousseeuw P.J., *Computing Depth Contours of Bivariate Point Clouds*, „Computational Statistics & Data Analysis" 1996, 23, s. 153–168.

Ruts I., Rousseeuw P.J., *Isodepth: A Program for Depth Contours*, w: *Proceedings in Computational Statistics*, Prat A. (red.), COMPSTAT, Physica, Heidelberg 1996.

Tukey J.W., *Mathematics and Picturing Data*, Proceedings of the 1974 international congress of mathematicians, R. James (red.), Vancouver 1975, vol. 2, s. 523–531.

Wielki J., *Elektroniczny marketing poprzez Internet*, Wydawnictwo Naukowe PWN, Warszawa–Wrocław 2000.

Zuo Y., Serfling R., *General Notions of Statistical Depth Function*, „Annals of Statistics" 2000, 28, 461–482, URL, http://dx.doi.org/10.1214/aos/1016218226 (dostęp: 5.07.2018).

Źródła sieciowe

https://cran.r-project.org/web/packages/depth/depth.pdf (dostęp: 15.06.2018).

https://cran.r-project.org/web/packages/DepthProc/index.html (dostęp: 15.06.2018).

https://www.ecommerce-europe.eu/press-item/european-ecommerce-report-2017-released-ecommerce-continues-prosper-europe-markets-grow-different-speeds/#_ftnref1 (dostęp 10.07.2018).

http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do (dostęp 10.06.2018).

* * *

Depth contours and a scale curve in the assessment of e-commerce diversity in European Union countries

Abstract

This article presents the application of observation depth contours in a sample and a scale curve in a two-dimensional data analysis. Numerical data on e-commerce in EU countries was used to illustrate this application.

Keywords: Tukey's observation depth measure in a sample, depth contours