Małgorzata Kobylińska[1]

Koncepcja zanurzania obserwacji w próbie w statystycznej analizie danych dotyczących handlu elektronicznego w przedsiębiorstwach

1. Wstęp

Przeprowadzając proces badawczy, mamy na celu między innymi wszechstronne zbadanie zjawiska oraz wykrycie pewnych prawidłowości i związków zachodzących w badanej zbiorowości. Spośród wielu metod statystycznej analizy danych często wykorzystywane są te, które umożliwiają grupowanie obiektów względem wartości cech diagnostycznych. Wykorzystując odpowiednie kryterium klasyfikacji, uzyskujemy klasy obiektów podobnych do siebie ze względuna wartości badanych cech. Wielowymiarowa analiza statystyczna stała się ważnym narzędziem wykorzystywanym w celu klasyfikacji oraz porządkowania obiektów opisanych za pomocą kilku cech[2].

Wraz z rozwojem technologii informatycznych pojawiły się nowe możliwości dotyczące wykorzystania komputerów w statystycznej analizie danych. Dostępne pakiety komputerowe umożliwiają wykonanie coraz bardziej skomplikowanych analiz statystycznych dotyczących danych wielowymiarowych.

W artykule zaprezentowano użyteczność wybranych metod zanurzania obserwacji w próbie w statystycznej analizie danych.W tym celu zostały wykorzystane dane liczbowe dotyczące handlu elektronicznego w przedsiębiorstwach w województwach Polski. Do obliczeń posłużono się pakietami komputerowymi środowiska R, które umożliwiły wyznaczenie miar zanurzania oraz wykonanie wykresów konturów zanurzania obserwacji w próbie. Program R jest wykorzystywany do badań naukowych oraz dydaktyki na licznych uczelniach na świecie[3].

---

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

[2] T. Panek, *Statystyczne metody wielowymiarowej analizy porównawczej*, Oficyna Wydawnicza SGH, Warszawa 2009; T. Grabiński, S. Wydymus, A. Zeliaś, *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, WN PWN, Warszawa 1989.

[3] M. Walesiak, G. Gatnar, *Statystyczna analiza danych z wykorzystaniem programu R*, WN PWN, Warszawa 2009.

2. Metody badawcze

Zagadnienia związane z zanurzaniem obserwacji zostały zapoczątkowane przez J.W. Tukey'a[4]. Koncepcja wykorzystania zanurzania obserwacji w próbie w statystycznej analizie danych stała się tematem licznych publikacji, między innymi takich autorów, jak D.L. Donoho i M. Gasko[5], M. Kobylińska i W. Wagner[6], D. Kosiorowski[7], R.Y. Liu i in.[8], P.J. Rousseeuw i I. Ruts[9].

Zanurzanie obserwacji w próbie może być narzędziem wykorzystywanym między innymi w celu porządkowania obserwacji wielowymiarowych względem „odstawania" od centrum próby lub do ich wizualizacji. Wykresy konturów zanurzania umożliwiają określenie centralnego skupienia zbiorów danych, wyznaczenie obserwacji nietypowych lub określenie symetrii i koncentracji rozkładu zmiennych. Poniżej zdefiniowane zostaną miary zanurzania obserwacji w próbie, które wykorzystywane zostały w pracy.

Niech $v\left[\Delta(x_1,x_2,...,x_{p+1})\right]$ będzie objętością p-wymiarowego sympleksu $\Delta(x_1,x_2,...,x_{p+1})$, którego wierzchołkami jest $p+1$ obserwacji z próby p-wymiarowej P_n^p o liczebności n. W przypadku p-wymiarowym liczba wszystkich możliwych sympleksów wynosi $N_{p+1}=\begin{pmatrix} n \\ p+1 \end{pmatrix}$, dla przypadku dwuwymiarowego jest równa $N_3=\begin{pmatrix} n \\ 3 \end{pmatrix}=\frac{1}{6}n\left(n-1\right)\left(n-2\right)$.

Miara zanurzania Oja[10] obserwacji θ w próbie p-wymiarowej P_n^p zdefiniowana jest jako:

---

[4] J.W. Tukey, *Mathematics and the Picturing of Data*, „Proceedings of the International Congress of Mathematicians" 1975, vol. 2.

[5] D.L. Donoho, M. Gasko,*Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness*,„The Annals of Statistics" 1992, vol. 20, no. 4.

[6] M. Kobylińska, W. Wagner, *Numerical aspects of determining measures and contours in depth for data in R^2*, „Acta Universitatis Lodziensis. Folia Oeconomica" 2002, t. 162.

[7] D. Kosiorowski, *Statystyczne funkcje głębi w odpornej analizie ekonomicznej*, Wydawnictwo Uniwersytetu Ekonomicznego, Kraków 2012.

[8] R.Y. Liu, J.M. Parelius, K. Singh, *Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference*, „The Annals of Statistics" 1999, vol. 27, no. 3, s. 783–858.

[9] P.J. Rousseeuw, I. Ruts, *Bivariate Location Depth*, „Applied Statistics" 1996, vol. 45, no. 4, s. 516–526.

[10] H. Oja, *Descriptive Statistics for Multivariate Distributions*, „Statistics of Probabilisty Letters" 1983, vol. 1, issue 6.

$$Ozan_p(\theta, P_n^p) = N_{p+1}^{-1}\left[1 + \sum_{1\le i_1\le i_2\le \ldots\le i_p\le n} v\left[\Delta(\theta, x_{i_1}, x_{i_2}, \ldots, x_{i_p})\right]\right], \tag{1}$$

gdzie: $v\left[\Delta(x_1, x_2, \ldots, x_{p+1})\right]$ określa objętość sympleksu wyznaczonego przez p punktów próby P_n^p oraz punkt θ.

Na podstawie powyższej definicji Y.J. Zuo i R. Serfling[11] zaproponowali miarę zanurzania Oja jako:

$$ZSzan_p(\theta, P_n^p) = N_{p+1}^{-1}\left[1 + \frac{\sum_{1\le i_1\le i_2\le \ldots\le i_p\le n} v\left[\Delta(\theta, x_{i_1}, x_{i_2}, \ldots, x_{i_p})\right]}{\sqrt{\det\left(P_n^p\right)}}\right]. \tag{2}$$

Dzięki wprowadzonej modyfikacji funkcja zanurzania Oja jest afinicznie niezmiennicza.

Miarą zanurzania Tukey'a ($Tzan_p$) punktu θ w próbie P_n^p nazywamy funkcję:

$$Tzan_p(\theta, P_n^p) = \frac{1}{n}\inf_{H_p}\left\{\theta \in H_p\right\}, \tag{3}$$

gdzie: H_p jest najmniejszą liczbą punktów badanej próby zawartą w zamkniętej półprzestrzeni w R^p, dla której linia graniczna przechodzi przez punkt θ[12].

Konturem zanurzania Tukey'a nazywamy zbiór $Con_k = \left\{\theta : zan_p(\theta, P_n^p) = k\right\}$ dla k=1,2,…[n/2], gdzie [n/2] jest częścią całkowitą liczby n/2. W tym przypadku zanurzanie obserwacji w próbie P_n^p jest rozumiane jako stopień konturu, do którego dana obserwacja należy. Wykresy konturów zanurzania przedstawiają wielokąty wypukłe, których wierzchołki wyznaczone zostały przed punkty przecięcia prostych rozdzielających półprzestrzeni H_p przechodzących przez punkt θ oraz inny punkt próby P_n^p. Punkt θ może być dowolnym punktem przestrzeni R^p lub punktem należącym do próby P_n^p[13].

W artykule do wyznaczenia miar zanurzania oraz sporządzenia odpowiednich wykresów wykorzystane zostały pakiety środowiska R: „DepthProc" autorstwa

[11] Y.J. Zuo, R. Serfling, *General notions of statistical depth function*, „The Annals of Statistics" 2000, vol. 28.

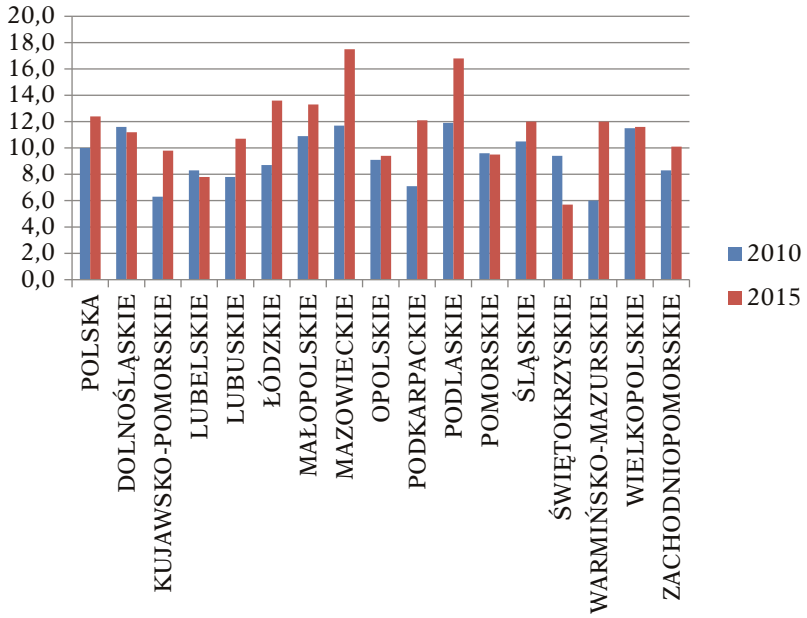
[12] R.Y. Liu, J.M. Parelius, K. Singh, op. cit.

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

D. Kosiorowskiego, M. Bociana, A. Wegrzynowskiej i Z. Zawadzkiego[14], „depth" autorstwa M. Genest, J.-C. Masse, J.-F. Plante[15] oraz „ddalpha" autorstwa O. Pokotylo, P. Mozharovskyi, R. Dyckerhoff, S. Nagy[16].

3. Analiza danych

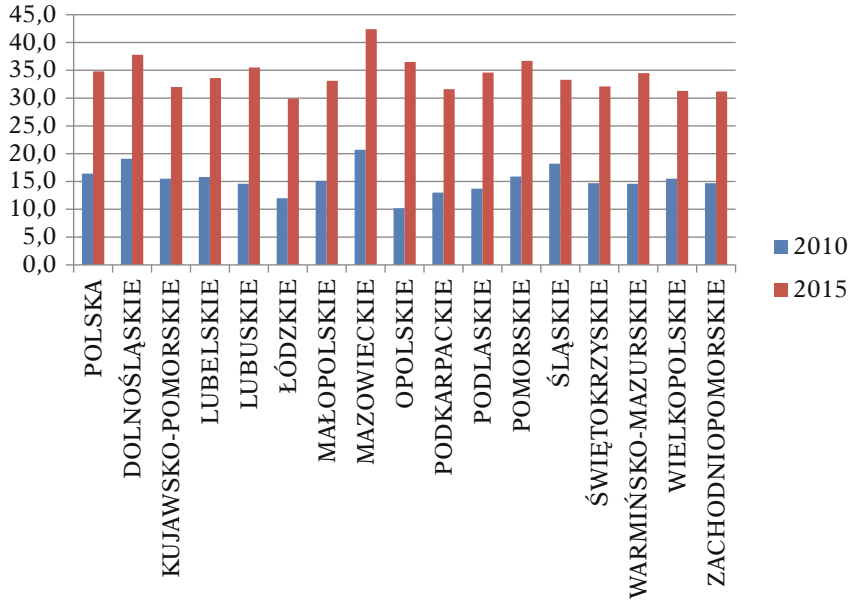
Wartości zmiennych diagnostycznych zaczerpniętoz Banku Danych Lokalnych GUS. Są to dane ilościowe dotyczące odsetka przedsiębiorstwsektora niefinansowego, zatrudniających więcej niż dziewięć osób, otrzymujących zamówienia (X1) oraz składających zamówienia (X2) poprzez sieci komputerowe w2010 i 2015 roku. Wartości analizowanych zmiennych zostały zaprezentowane na rysunkach 1 i 2.



Rysunek 1. Odsetek przedsiębiorstw otrzymujących zamówienia przez sieci komputerowe w 2010 oraz 2015 roku

Źródło: opracowanie własne na podstawie danych GUS.

[14] https://cran.r-project.org/web/packages/DepthProc/index.html (dostęp: 15.08.2017).
[15] https://cran.r-project.org/web/packages/depth/index.html (dostęp: 15.08.2017).
[16] https://cran.r-project.org/web/packages/ddalpha/index.html (dostęp: 15.08.2017).



Rysunek 2. Odsetek przedsiębiorstw składających zamówienia przez sieci komputerowe w 2010 i 2015 roku

Źródło: opracowanie własne na podstawie danych GUS.

Zauważyć można w badanych latach wzrost zainteresowania handlem elektronicznym wśród przedsiębiorstw. Najczęściej składały zamówienia przez sieć komputerową podmioty w województwie dolnośląskim oraz mazowieckim, w którymodsetek w 2015 roku wynosił 42,4%. Najniższymi wartościami tego wskaźnika w danych latach charakteryzowały się województwa opolskie (w 2010 r. – 10,2%) oraz łódzkie (w 2015 r. – 29,9%). W województwie opolskim w 2015 roku wartość tej zmiennej wzrosła najbardziej w porównaniu z 2010 rokiem, o 26,3 p.p. Odsetek przedsiębiorstw składających zamówienia zwykorzystaniem Internetu w 2015 roku w Polsce wynosił 34,13% i był o prawie 20 p.p. (18,92) wyższy w stosunku do roku 2010.

Mniejszym zainteresowaniem cieszyło się otrzymywanie zamówień z wykorzystaniem sieci komputerowych. Liderami były w danych latach odpowiednio województwo podlaskie (11,9% w roku 2010) oraz województwo mazowieckie (17,5% w roku 2015). W województwie świętokrzyskim w 2015 roku tylko co szóste przedsiębiorstwo (5,7%) otrzymywało zamówienia z wykorzystaniem sieci komputerowych. Wartość tego wskaźnika zmalała w tym województwie w porównaniu z 2010 rokiem o 3,7 p.p. Można zauważyć, że odsetek przedsiębiorstw, które korzystały z tej formy otrzymywania zamówień w 2015 roku, był

tylko o 2,16 p.p. wyższy w stosunku do roku 2010. Większe zróżnicowanie w ujęciu terytorialnym zaobserwować można w przypadku odsetka otrzymywanych zamówień. Współczynniki zmienności w tym przypadku wynosiły odpowiednio w latach 20,81% oraz 26,11%. Najmniejszym zróżnicowaniem charakteryzowały się województwa w 2015 roku ze względu na wartość zmiennej X2 (V=9,20%).

W tabeli 1 zamieszczono wartości miary zanurzania Tukey'a oraz simpleksowego Oja wyznaczone na podstawie wzorów 2 i 3. Poszczególnym województwom przyporządkowane zostały rangi zgodnie z odpowiadającymi im wartościami tych miar. Rangę 1 przypisano wartości najmniejszej. Województwom, którym odpowiada w danych latach wartość miary zanurzania Tukey'a równa zero, zostały przypisane rangi odpowiednio 3,5 i 3,0. Województwa te są wierzchołkami powłok wypukłych zbiorów danych. W 2010 roku powłoka wypukła została wyznaczona przez sześć województw (kujawsko-pomorskie, mazowieckie, opolskie, podkarpackie, podlaskie, warmińsko-mazurskie), natomiast w 2015 roku przez pięć województw (łódzkie, mazowieckie, podlaskie, pomorskie, świętokrzyskie). Wierzchołki powłoki wypukłej tworzą województwa, których zmienne diagnostyczne przyjmują niskie lub wysokie wartości. Województwa mazowieckie oraz podlaskie należą do powłoki wypukłej w każdym badanym lat. Odsetek przedsiębiorstw, które otrzymywały zamówienia przez sieci komputerowe, był w tym przypadku wyższy od przeciętnej w kraju. Województwo mazowieckie należy do powłoki wypukłej ze względu na znacznie wyższe wartości wszystkich analizowanych wskaźników w badanych latach w porównaniu ze średnią w Polsce.

Wartości miary zanurzania simpleksowego Oja pozwoliły na wyznaczanie dwuwymiarowych wektorów medianowych. Odpowiadają im województwa z najwyższymi wartościami tej miary. W kolejnych latach są to województwa warmińsko-mazurskie oraz wielkopolskie. Najmniejsza wartość miary zanurzania Oja w 2015 roku odpowiada województwu łódzkiemu, w którym zanotowano najniższy odsetek przedsiębiorstw składających zamówienia z wykorzystaniem sieci komputerowych. Wartość wskaźnika nie przekroczyła w tym przypadku 30% i uplasowała się o 4,23 p.p. poniżej średniej w kraju.

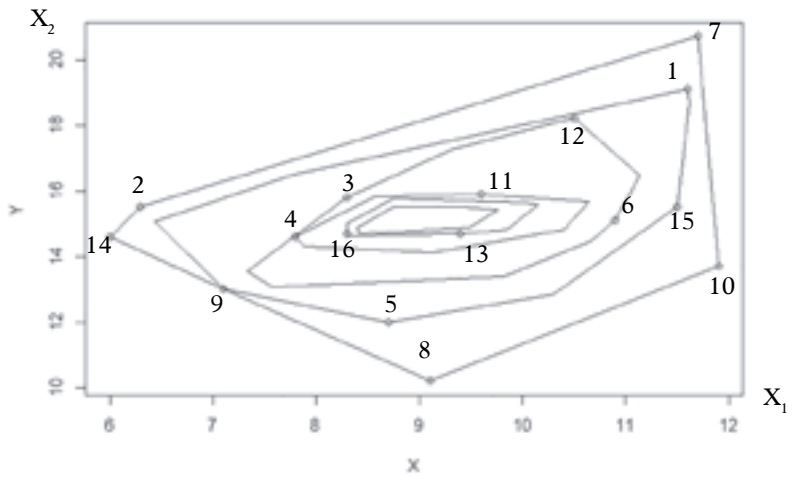
Wartości median wyznaczone dla miar zanurzania Oja w danych latach (0,5790, 0,5516) pozwoliły na wyodrębnienie zbiorów zawierających 50% województw, które położone są najbardziej centralnie w danych zbiorach danych. Województwa lubelskie, lubuskie, śląskie oraz warmińsko-mazurskie należą do tych zbioru w każdym z badanych lat.

Tabela 1. Miary zanurzania Tukey'a oraz simpleksowego Oja

Lp.	Województwo	Tzan $_{2010}$	Ozan$_{2010}$	R_{T_2010}	R_{O_2010}	Tzan $_{2015}$	Ozan$_{2015}$	R_{T_2015}	R_{O_2015}
1.	Dolnośląskie	0,0625	0,5855	8	10	0,0625	0,5170	8,5	7
2.	Kujawsko-Pomorskie	0,0000	0,5195	3,5	5	0,1250	0,5024	12,5	5
3.	Lubelskie	0,1250	0,7465	11	15	0,0625	0,5922	8,5	10
4.	Lubuskie	0,1875	0,6845	13,5	14	0,2500	0,6322	15	13
5.	Łódzkie	0,0625	0,5792	8	9	0,0000	0,3771	3	1
6.	Małopolskie	0,1250	0,4837	11	2	0,1250	0,4892	12,5	3
7.	Mazowieckie	0,0000	0,5787	3,5	8	0,0000	0,4123	3	2
8.	Opolskie	0,0000	0,4845	3,5	3	0,0625	0,6110	8,5	11
9.	Podkarpackie	0,0000	0,5426	3,5	7	0,0625	0,7643	8,5	14
10.	Podlaskie	0,0000	0,4712	3,5	1	0,0000	0,5335	3	8
11.	Pomorskie	0,1875	0,5285	13,5	6	0,0000	0,6155	3	12
12.	Śląskie	0,1250	0,6100	11	12	0,3125	0,5696	16	9
13.	Świętokrzyskie	0,2500	0,6057	15,5	11	0,0000	0,4904	3	4
14.	Warmińsko-Mazurskie	0,0000	0,8031	3,5	16	0,1875	0,7779	14	15
15.	Wielkopolskie	0,0625	0,4952	8	4	0,0625	0,7954	8,5	16
16.	Zachodniopomorskie	0,2500	0,6598	15,5	13	0,0625	0,5089	8,5	6

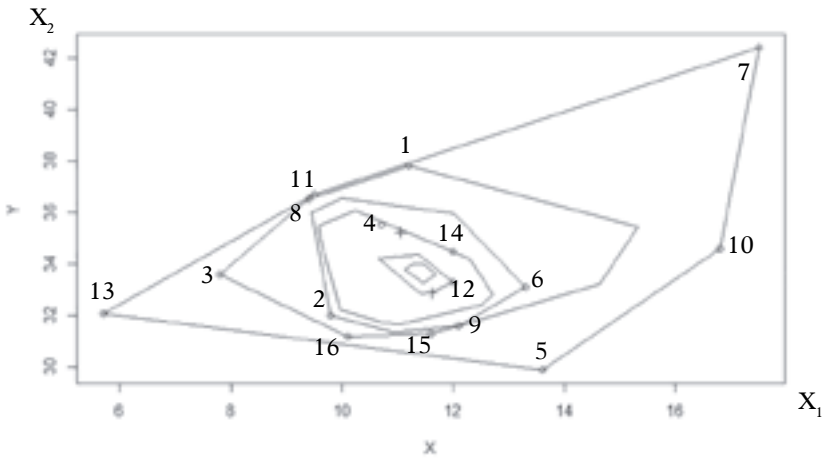
Źródło: opracowanie własne na podstawie obliczeń w pakietach „DepthProc" i „ddalpha".

Na rysunkach 3–6 przedstawiono wykresy konturów zanurzania Tukey'a na płaszczyźnie dwuwymiarowej oraz wykresy perspektywiczne tych konturów. Powłoki wypukłe są najmniejszymi wielokątami wypukłymi zawierającymi wszystkie obserwacje zbiorów danych. W 2015 roku na kształt powłoki wypukłej wpłynęły wartości zmiennych województwa mazowieckiego oraz świętokrzyskiego. Odpowiednio ze względu na wysokie wartości zmiennych (województwo mazowieckie) oraz w przypadku województwa świętokrzyskiego najniższą wartość wskaźnika dotyczącego otrzymywania zamówień przez sieci komputerowe (5,7%). Wartość tego wskaźnika w tym przypadku była prawie trzykrotnie niższa w porównaniu z województwem mazowieckim. Wykresy konturów zanurzania Tukey'a umożliwiły graficzne przedstawienie koncentracji oraz siły i kierunku zależności pomiędzy rozważanymi danymi. Kształt konturów zanurzania wskazuje na korelację dodatnią pomiędzy rozważanymi zmiennymi w danych latach, przy czym silniejsza jest ona w roku 2010. Współczynniki korelacji liniowej Pearsona wynoszą odpowiednio $r_{2010} = 0,43, r_{2015} = 0,34$.



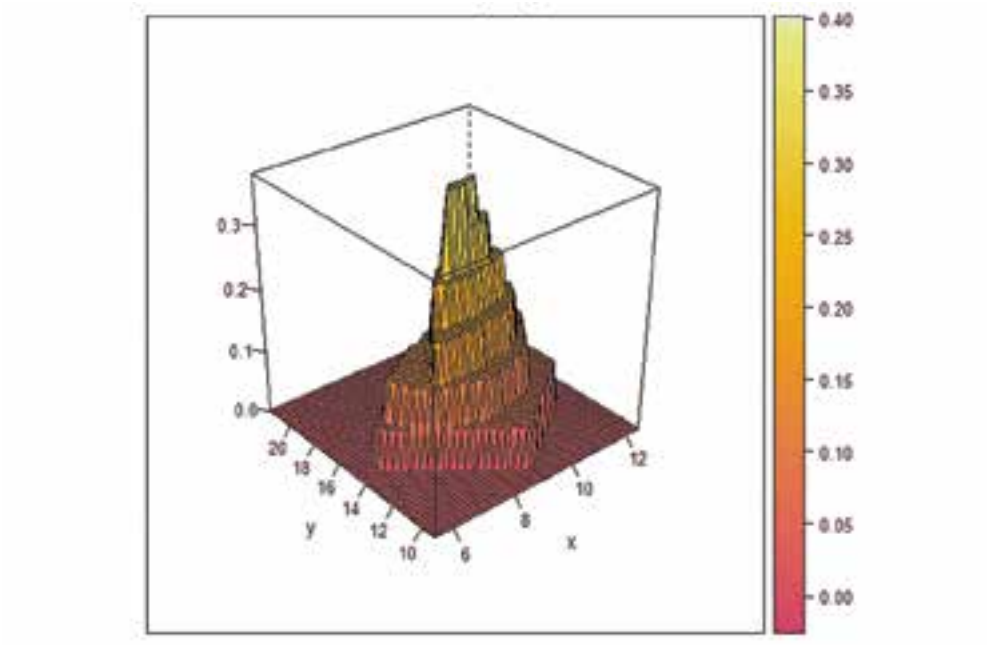
Rysunek 3. Wykres konturów zanurzania Tukey'a dla danych z 2010 roku

Źródło: opracowanie własne z wykorzystaniem pakietu „depth".



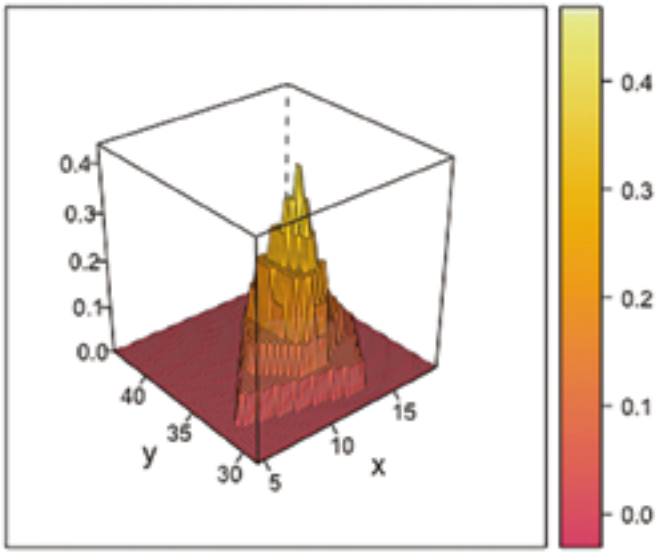
Rysunek 4. Wykres konturów zanurzania Tukey'a dla danych z 2015 roku

Źródło: opracowanie własne z wykorzystaniem pakietu „depth".



Rysunek 5. Wykres perspektywiczny dla danych z 2010 roku

Źródło: opracowanie własne z wykorzystaniem pakietu „DepthProc".



Rysunek 6. Wykres perspektywiczny dla danych z 2015 roku
Źródło: opracowanie własne z wykorzystaniem pakietu „DepthProc".

4. Podsumowanie

Przeprowadzona analiza wskazuje obszar zastosowań wybranych metod opartych na koncepcji zanurzania obserwacji w próbie. Korzystając z wartości miary zanurzania, uzyskano rangowanie województw względem oddalenia od centralnego skupienia (mediany zanurzania) oraz wyznaczono jednostki, które mogą być uznane na „odstające" ze względu na wartości badanych cech. Województwa, które ze względu na wartość miary zanurzania położone są najbardziej centralnie w zbiorach danych, można uznać za „typowe" ze względu na wartości dotyczące odsetka przedsiębiorstw otrzymujących oraz składających zamówienia przez sieci komputerowe. Wykresy konturów zanurzania pozwoliły na wizualizację danych oraz na zobrazowanie pewnych własności rozpatrywanych zbiorów.

Metody analizy danych oparte na zanurzaniu obserwacji w próbie mogą znaleźć szersze zastosowanie w statystycznej analizie danych oraz mogą stanowić uzupełnienie klasycznych metod wielowymiarowej analizy statystycznej.

Bibliografia

Donoho D.L., Gasko M., *Breakdown Properties of Location Estimates Based on Half-space Depth and Projected Outlyingness*,„The Annals of Statistics"1992, vol. 20, no. 4, s. 1803–1827.

Grabiński T., Wydymus S., Zeliaś A., *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, WN PWN, Warszawa 1989.

Kobylińska M., Wagner W.,*Numerical aspects of determining measures and contours in depth for data in* R^2, „Acta Universitatis Lodziensis. Folia Oeconomica" 2002, t. 162, s. 19–32.

Kosiorowski D., *Statystyczne funkcje głębi w odpornej analizie ekonomicznej*, Wydawnictwo Uniwersytetu Ekonomicznego, Kraków 2012.

Liu R.Y., Parelius J.M., Singh K.,*Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference*, „The Annals of Statistics" 1999, vol. 27, no. 3, s. 783–858.

Oja H., *Descriptive Statistics for Multivariate Distributions*, „Statistics of Probabilisty Letters" 1983, vol. 1, issue 6, s. 327–323.

Panek T., *Statystyczne metody wielowymiarowej analizy porównawczej*, Oficyna Wydawnicza SGH, Warszawa 2009.

Rousseeuw P.J., Ruts I., *Bivariate Location Depth*, „Applied Statistics"1996, vol. 45, no. 4, s. 516–526.

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

Tukey J.W., *Mathematics and the Picturing of Data*, „Proceedings of the International Congress of Mathematicians" 1975, vol. 2, s. 523–531.

Walesiak M., Gatnar G., *Statystyczna analiza danych z wykorzystaniem programu R*, WN PWN, Warszawa 2009.

Zuo Y.J., Serfling R., *General notions of statistical depth function*, „The Annals of Statistics 2000", vol. 28, s. 461–482.

Źródła sieciowe

https://cran.r-project.org/web/packages/DepthProc/index.html (dostęp: 15.08.2017).

https://cran.r-project.org/web/packages/depth/index.html (dostęp: 15.08.2017).

https://cran.r-project.org/web/packages/ddalpha/index.html (dostęp: 15.08.2017).

https://bdl.stat.gov.pl/BDL/dane/podgrup/tablica (dostęp: 10.07.2017).

(dane dotyczące odsetka przedsiębiorstw otrzymujących zamówienia oraz składających zamówienia z wykorzystaniem sieci komputerowych).

* * *

Concept of Observation Depth Measure in the Statistical Analysis of E-Commerce Data in Enterprises

Summary

This article presents the application of selected methods based on the observation depth measure in statistical data analysis. The figures concerning e-commerce among the enterprises of the Polish provinces in 2010 and 2015 were used.

Keywords: multi-dimensional data analysis, observation depth measure, depth contour, depth median.