

Arkadiusz Maciuk

Akademia Ekonomiczna we Wrocławiu

GDZIE JEST ŚRODEK POLSKI, CZYLI O PROBLEMIE WYBORU REPREZENTANTA ZBIORU

Najczęściej stosowanymi reprezentantami zbioru są średnia arytmetyczna, mediana i modalna zbioru. Średnia arytmetyczna szczególnie ważną funkcję pełni w statystyce i w zagadnieniach pokrewnych, takich jak teoria pomiaru czy ekonometria. Większość rozkładów probabilistycznych ma tę własność, iż są skoncentrowane wokół pewnego punktu, a prawdopodobieństwo wystąpienia elementu o takim rozkładzie maleje bardzo szybko w miarę oddalania się od tego „centralnego” punktu. Tę własność wykorzystuje metoda najmniejszych kwadratów, która zakłada, że spadek ten jest co najmniej proporcjonalny do kwadratu odległości od tego centrum. Estymatorem takiego „centrum rozkładu probabilistycznego” w metodzie najmniejszych kwadratów jest właśnie średnia arytmetyczna [Brandt 1998]. Jeśli jednak spadek ten jest wolniejszy, lepszym estymatorem jest mediana. Na przykład jeśli w badanym zbiorze znajdują się elementy nietypowe, znacznie różniące się (znacznie oddalone) od większości innych elementów, ich wpływ na średnią arytmetyczną może być nieproporcjonalnie duży w stosunku do ich (relatywnej) ważności. Jeśli istnieje podejrzenie, że tego typu sytuacja ma miejsce, zazwyczaj za reprezentanta zbioru przyjmuje się właśnie medianę zbioru, która jest mniej podatna na tego typu „elementy skrajne”.

Oprócz średniej arytmetycznej i mediany, trzecim, stosunkowo często stosowanym reprezentantem zbioru jest modalna (moda), czyli element najczęściej występujący (bądź inaczej – element o największej wadze). Ma on tę cechę, iż nie zależy od kształtu analizowanego zbioru, tylko od jego miary.

Wspólną cechą tych trzech reprezentantów jest to, że do ich wyznaczenia nie są potrzebne skomplikowane algorytmy – co w przypadku zbiorów dużych i wielowymiarowych jest bardzo wygodne. Ta wygoda sprawia, że inne możliwości wy-

boru reprezentanta są rzadko stosowane, a przez to mało znane. Po co szukać innych, „udziwnionych” reprezentantów, skoro do dyspozycji mamy gotowe, proste wzory, dające zadowalające rozwiązanie? Jednak takie podejście jest niepoprawne. Wybór reprezentanta powinien być poprzedzony ustaleniem kryteriów wyboru – uwarunkowanych m.in. rodzajem zagadnienia. Niewielka zmiana kryteriów może powodować, że nasz wybór będzie znacznie odmienny. Tymczasem wygoda w postaci istnienia gotowych i nieskomplikowanych wzorów sprawia, że często stosuje się podejście zupełnie inne. Dokonuje się wyboru reprezentanta bez przyjmowania założeń uzasadniających ten wybór bądź automatycznie przyjmuje się pewne założenia nie mające żadnego związku z analizowanym zagadnieniem.

Przykładem takiego postępowania może być właśnie wybór średniej arytmetycznej na reprezentanta zbioru w przypadku zagadnień nieprobabilistycznych, takich jak szukanie optymalnej lokalizacji. Jeśli mamy zbiór pomiarów pewnej nieznannej wartości, różniących się wzajemnie pomiędzy sobą, naturalnym założeniem jest przyjęcie, że pomiary te obarczone są pewnym błędem, mającym rozkład normalny o średniej zero i pewnej stałej wariancji. Przy takim założeniu szukaną wartość najlepiej estymuje, w sensie metody najmniejszych kwadratów, średnia arytmetyczna (zob. [Brandt 1998]). Innymi słowy, założenie, że dany zbiór danych ma rozkład normalny, uzasadnia wybór średniej arytmetycznej na jego reprezentanta.

Jednak jeśli czynnik losowy w zagadnieniu nie występuje (jak w przypadku problemów związanych z lokalizacją działalności gospodarczej), uzasadnienie to traci sens. Wybór jako reprezentanta średniej arytmetycznej wymaga wtedy innego uzasadnienia – czego w podobnych przypadkach praktycznie się nie czyni. Co więcej, w większości tego typu zagadnień nie można zakładać, że istotnie wielkość (ważność, miara) punktów maleje w miarę zwiększania się ich odległości od jednego, centralnego punktu. Stąd stosowanie średniej arytmetycznej bądź mediany nie jest ani korzystne, ani merytorycznie poprawne.

Czy istnieje alternatywa? Oczywiście możliwości wyboru jest bardzo dużo. W zagadnieniach ekonomicznych związanych z lokalizacją większe zastosowanie mają reprezentanci o charakterze „lokalnym”, czyli tacy, którzy są bardziej zależni od miary niż od kształtu analizowanego zbioru.

Niech (X, w) będzie pewnym danym, ograniczonym zbiorem punktów X z m -wymiarowej przestrzeni liczb rzeczywistych $\mathbf{R}^m$ ($m < \infty$) z miarą w ($w(X) < \infty$), k -dodatnim parametrem, a $d(x, p)$ – odległością pomiędzy ustalonym punktem x ze zbioru X , a punktem $p \in \mathbf{R}^m$.

Definicja 1. Reprezentant zbioru (X, w) to punkt realizujący minimum funkcji

$$f_{X,w}(p) = f_{X,w}(p; k, d) = \left(\sum_{x \in X} d(x, p)^k w_x \right)^{1/k} = \left(\sum_{i=1}^n d(x_i, p)^k w_i \right)^{1/k}, \quad (1)$$

gdzie $w_x = w_i$ – miara punktu x ¹.

Tak zdefiniowany reprezentant ma naturalną interpretację. Im bardziej rozproszone są punkty ze zbioru X , tym większą wartość ma funkcja (1). Punkt, w którym ta funkcja ma wartość najmniejszą, jest punktem, który minimalizuje łączną odległość do wszystkich punktów ze zbioru X – a więc punktem, który najlepiej reprezentuje ten zbiór.

W takim ujęciu reprezentant zbioru zależy nie tylko od zbioru X i jego miary w , ale również od wyboru metryki d i wartości parametru k . Parametr k można określić mianem parametru kształtu zbioru, gdyż im większa jego wartość, tym większą rolę odgrywa kształt analizowanego zbioru. Jeśli $k = 2$, wpływ daleko położonych elementów na potencjalnego reprezentanta rośnie proporcjonalnie do kwadratu ich odległości od niego. W skrajnym przypadku, gdy $k \rightarrow \infty$, reprezentant w ogóle nie zależy od miary zbioru, tylko od jego elementów brzegowych, gdyż

$$f_{X,w}(p; d, \infty) = \lim_{k \rightarrow \infty} \left(\sum_{x \in X} d(x, p)^k w_x \right)^{1/k} = \max_{x \in X} d(x, p)^2. \text{ Stąd minimum funkcji } f$$

względem p realizuje punkt będący środkiem okręgu opisującego analizowany zbiór. Punkt ten będziemy nazywać środkiem Czebyszewa, w analogii do nazwy metryki Czebyszewa d_∞ .

W większości przypadków znalezienie minimum funkcji (1) jest skomplikowane ze względów natury obliczeniowej. Jednak w pewnych szczególnych przypadkach jest względnie proste. Zauważmy, iż jeśli k jest liczbą skończoną, wzór (1) można uprościć, pomijając symbol pierwiastka (operacja potęgowania funkcji nie zmienia jej minimum). Gdy za odległość d przyjmiemy metrykę Minkowskiego

$$d_t(x, p) = \left(\sum_{i=1}^m |x_i - p_i|^t \right)^{1/t}, \text{ gdzie } k = t, \text{ wtedy zagadnienie znalezienia minimum w}$$

zbiorze z $\mathbf{R}^m$ sprowadza się do znalezienia m minimów w m zbiorach z przestrzeni $\mathbf{R}^3$.

¹ Wzór ten jest szczególnym przypadkiem wzoru $f_{X,w}(p) = \left(\int d(x, p)^k dw(x) \right)^{1/k}$. Gdy miara w

jest miarą singularną, skupioną w punktach, przybiera on postać (1). Postać ogólna jest niepraktyczna ze względów rachunkowych, dlatego też wygodniej jest rozpatrywać pierwszą, dyskretną postać. Właściwie nie tracimy na ogólności, zawiązując nasze rozważania do tego przypadku, gdyż w większości zagadnień mamy do czynienia ze zbiorami dyskretnymi; a co ważniejsze, każdą miarę ciągłą można przybliżać, z dowolnie małym błędem, miarą dyskretną [Billingsley 1987].

² Wzór ten jest prostym uogólnieniem prostej zależności. Jeśli $a_1, \dots, a_k \geq 0$ i $w_1, \dots, w_k \geq 0$, to $\lim_{n \rightarrow \infty} \sqrt[n]{a_1^n w_1 + \dots + a_k^n w_k} = \max \{a_1, \dots, a_k\}$.

³ $f^k_{X,w}(p; d_k, k) = \sum_{x \in X} (d_k(x, p))^k w_x = \sum_{i=1}^n \sum_{x \in X} |x_i - p_i|^k w_x$.

Stąd $\min_p f^k_{X,w}(p; d_k, k) = \sum_{i=1}^n \min_{p_i} \sum_{x \in X} |x_i - p_i|^k w_x$.

Na przykład jeśli za odległość przyjmiemy metrykę euklidesową d_2 , a $k = 2$, minimum funkcji (1) realizuje średnia arytmetyczna zbioru⁴. Gdy zaś za odległość przyjmiemy metrykę miejską d_1 , a $k = 1$, minimum funkcji (1) realizuje mediana zbioru X ⁵ (zob. tab. 1).

Tabela 1. Przykłady reprezentantów zbioru (według definicji 1)

Nazwa reprezentanta	k	Metryka
Środek Czebyszewa	∞	dowolna
Mediana	1	d_1
Średnia arytmetyczna	2	d_2
Modalna	dowolne	metryka dyskretna

Źródło: opracowanie własne.

Jednak już w sytuacji, gdy odległość jest odległością euklidesową d_2 , a $k = 1$, nie istnieje wzór pozwalający bezpośrednio wyznaczyć minimum funkcji (1). Mimo wszystko znalezienie takiego minimum nie jest trudne. Można zastosować prosty algorytm polegający na wyznaczaniu kolejnych iteracji coraz bliższych poszukiwanemu minimum. Na początku wybiera się pewien punkt (na przykład średnią arytmetyczną). Dalej sprawdza się, jak szybko, wychodząc od tego punktu, funkcja maleje na różnych kierunkach. W tym celu wystarczy policzyć gradient funkcji i zbadać, w jakim kierunku jest najmniejszy (najbardziej ujemny). Jeśli w żadnym kierunku funkcja nie maleje (wszystkie pochodne cząstkowe są dodatnie), punkt ten jest szukanym minimum – co kończy algorytm. Jeśli nie, procedurę powtarza się, szukając minimum funkcji (1) na kierunku największego spadku. Za punkt początkowy podstawia się punkt realizujący to minimum. Następnie ponownie sprawdza się, wychodząc od nowego punktu, w którą stronę funkcja maleje najszybciej. Procedurę powtarza się dopóty, dopóki nie znajdzie się punktu będącego lokalnym minimum. Jeśli funkcja (1) jest funkcją wypukłą, takie lokalne minimum jest jednocześnie jej minimum globalnym. Algorytm ten jest stosunkowo prosty i daje poprawne rozwiązanie już po kilku iteracjach.

Jednak w sytuacji gdy ta funkcja nie jest funkcją wypukłą, zawodzi, gdyż wtedy może mieć minima lokalne. Można posłużyć się w tym miejscu analogią z piłeczką i dołem. Puszczona w dół piłeczka będzie opadać, póki nie zatrzyma się

$$^4 \text{ Średnia arytmetyczna to punkt } \bar{x}_w = \left(\frac{\sum_{i=1}^n x_{1i} w_i}{\sum_{i=1}^n w_i}, \frac{\sum_{i=1}^n x_{2i} w_i}{\sum_{i=1}^n w_i}, \dots, \frac{\sum_{i=1}^n x_{mi} w_i}{\sum_{i=1}^n w_i} \right)^T = \frac{1}{w(X)} \sum_{x \in X} x w_x$$

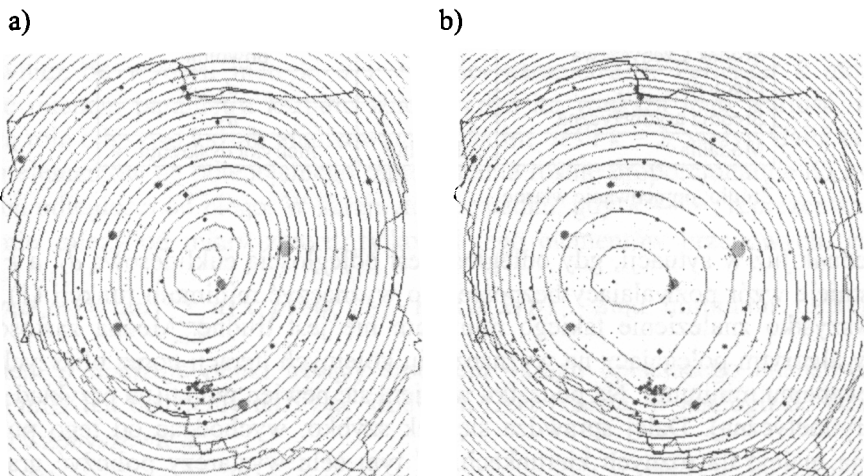
(zob. [Brandt 1998; Jajuga 1999]).

⁵ Medianą j -tych współrzędnych jest liczba $\hat{x}_{j,w} = x_{j|s|}$, gdzie s – indeks taki, że

$$\sum_{i=1}^{s-1} w_{|i|} < \frac{1}{2} \sum_{i=1}^n w_i \leq \sum_{i=1}^s w_{|i|}. \text{ Medianą zbioru } X \text{ jest punkt } \hat{x}_w = (\hat{x}_{1,w}, \dots, \hat{x}_{m,w})^T \text{ [Brandt 1998].}$$

w najniższym punkcie dołu. Jeśli jednak dół składa się z kilku dołków różnej głębokości, piłka zatrzyma się w pierwszym, przypadkowym dołku leżącym na jej drodze, niekoniecznie w tym najgłębszym.

Powstaje pytanie, czy funkcja (1) zawsze jest funkcją wypukłą? Jest taka np. wtedy, gdy $k \geq 1$ i odległość $d(x, p)$, traktowana jako funkcja względem zmiennej p , jest funkcją wypukłą⁶, (jak metryka Minkowskiego – zob. rys. 1).



Rys. 1. Warstwy funkcji (1) przy $d := d_2$ i $k = 1$ (a) oraz $k = 2$ (b)

Źródło: opracowanie własne.

Jednak w wielu przypadkach funkcja (1) nie jest wypukła. Wystarczy, że za odległość d przyjmiemy odległość dyskretną albo odległość określoną za pomocą wzoru

$$d'_r(x, p) = \frac{d(x, p)}{r + d(x, p)} \quad (2)$$

lub

$$d''_r(x, p) = 1 - 2^{-d(x, p)/r}, \quad (3)$$

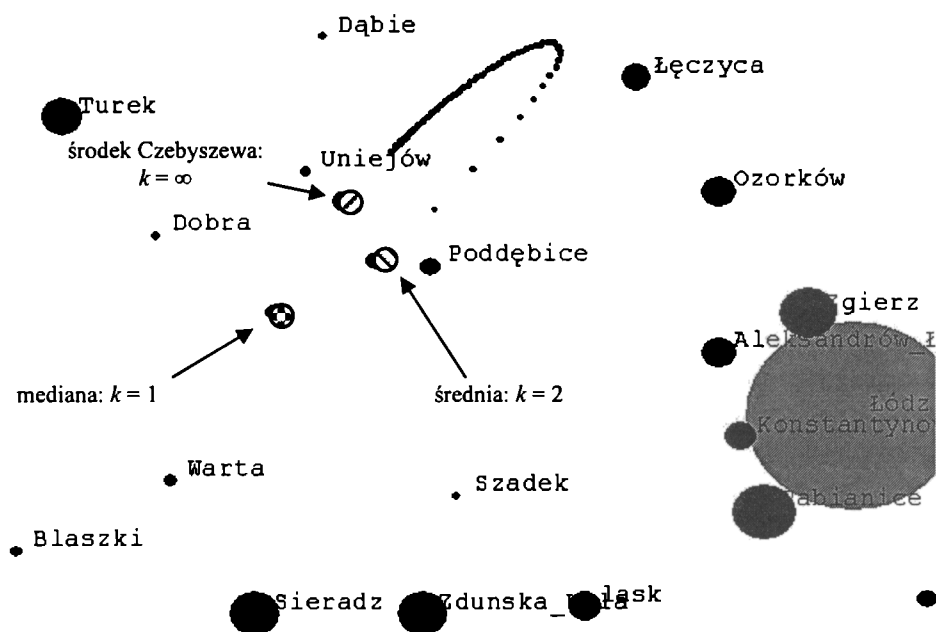
gdzie r jest dowolnym dodatnim parametrem, a d pewną, dowolną metryką – chociażby metryką Minkowskiego.

W przypadku metryki dyskretnej (i gdy k jest dowolnym skończonym parametrem) reprezentantem jest modalna zbioru – reprezentant stosunkowo łatwy do wyznaczenia. Jednak gdy za odległość przyjmiemy dowolną metrykę określoną wzorem (2) lub (3), znalezienie minimum funkcji (1) może być o wiele trudniejsze, gdyż może wymagać wielu żmudnych obliczeń (polegających na obliczaniu wartości funkcji (1) nawet na całym analizowanym obszarze).

⁶ Dowolna kombinacja liniowa o nieujemnych współczynnikach nieujemnych funkcji wypukłych jest funkcją wypukłą (zob. [Maciuk 1999]).

W dobie powszechnego dostępu do komputerów trudności związane z „pracochłonnością” algorytmów są coraz mniejsze. Czy jednak istnieje potrzeba wyznaczania takich reprezentantów? W dużym stopniu zależy to od rodzaju i specyfiki analizowanego zagadnienia. Gdy wpływ elementów skrajnych jest niewielki, czy nawet pomijalny, wówczas tego typu reprezentant lepiej się sprawdza niż reprezentant klasyczny, jak średnia czy mediana. Szczególnie ważne jest to w zagadnieniach związanych z lokalizacją podmiotów gospodarczych, jak fabryka, magazyn, centrum logistyczne itp.

Załóżmy, że zbiór X jest zbiorem potencjalnych klientów, w – miarą ich wielkości (ważności). Punkt p oznaczać będzie lokalizację ośrodka, który ma ich obsługiwać. Im większa odległość pomiędzy parą punktów x a p , tym większa strata związana z obsługą klienta x przez ośrodek w punkcie p . Wartość funkcji (1) w punkcie p oznaczać będzie łączną stratę z obsługi wszystkich punktów ze zbioru X (przy uwzględnieniu ich wielkości w_x) przez punkt p . Znalezienie punktu realizującego minimum tej funkcji równoważne jest więc z wyznaczeniem punktu, z którego dotarcie do wszystkich punktów $x \in X$ będzie powodowało straty najmniejsze z możliwych.



Rys. 2. Przykłady reprezentantów zbioru gmin Polski (według definicji 1) z liczbą ludności jako miarą przy założeniach, iż $d := d_2$ i przy $k = 1, 2, \dots$

Źródło: opracowanie własne.

W tego typu sytuacjach kwestią zasadniczą jest ocena zależności pomiędzy wzrostem straty na obsłudze pojedynczego klienta x w miarę wzrostu odległości od ośrodka p . Czy taka strata jest proporcjonalna do kwadratu odległości, w związku z czym uzasadnione jest wyznaczenie na reprezentanta zbioru średniej arytmetycznej? Czy może spadek ten jest „tylko” proporcjonalny do odległości (reprezentant – mediana), a może jest dużo wolniejszy?

Zagadnienie to dobrze ilustruje następujący przykład. Niech zbiorem X będzie zbiór gmin Polski z liczbą mieszkańców jako jego miarą. Reprezentanta takiego zbioru można określić mianem środka Polski. Gdzie znajduje się taki środek?

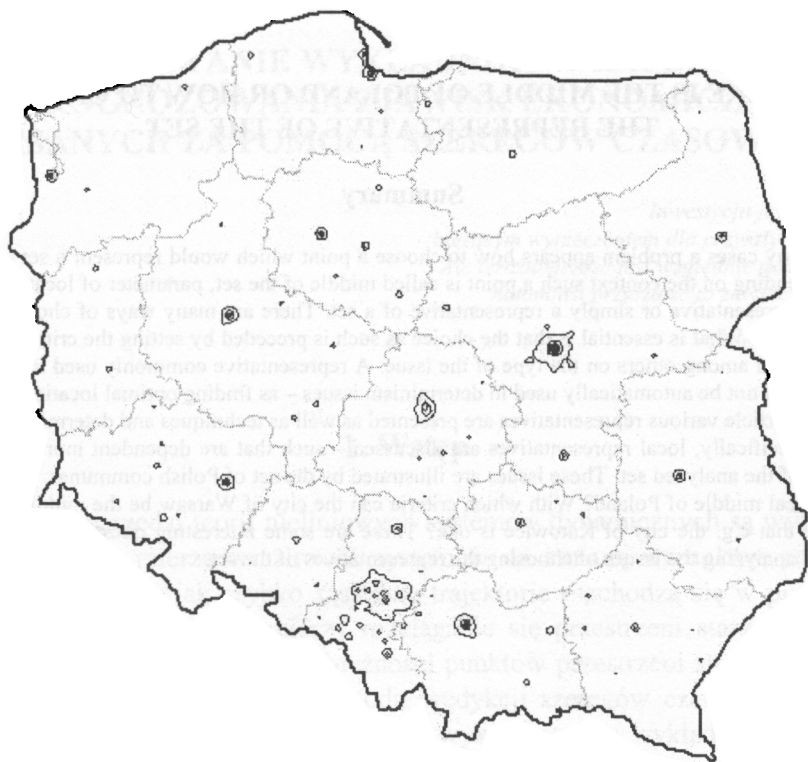
Według definicji 1 środek ten zależy od wyboru metryki d i wartości parametru kształtu k . Rysunek 2 przedstawia punkty realizujące minimum funkcji (1) przy $d = d_2$ i przy kolejnych wartościach parametru k ($k = 1, 2, \dots$). Ciekawym rezultatem jest to, iż wszystkie tak określone reprezentanty są położone stosunkowo blisko siebie, ok. 50 km na zachód od Łodzi, pomiędzy miejscowościami Dobra i Poddębice.

Jakie jest ekonomiczne znaczenie tego typu „środków”? Dla przeciętnego przedsiębiorcy niewielkie, gdyż jest on zainteresowany nie tyle minimalizacją łącznej odległości od „wszystkich”, co wybraniem na miejsce swojej działalności punktu, w pobliżu którego znajduje się jak największa liczba potencjalnych klientów, dostawców itd. Z jego punktu widzenia lepszym reprezentantem (czyli lepszym miejscem na prowadzenie działalności gospodarczej) jest modalna (czyli miasto Warszawa) bądź też reprezentant określony przez odległość daną wzorem (2) lub (3). Odległość Minkowskiego, traktowana jako funkcja względem p , wzrasta bez ograniczeń w miarę oddalania się od punktu x . Natomiast odległość dyskretna czy odległości określone wzorem (2) lub (3) mają tę własność, że od pewnego miejsca rosną stosunkowo wolno i nie przekraczają wartości 1. Można powiedzieć, że dla większości przedsiębiorców odległość we wzorze (1) nie jest spostrzegana jako funkcja liniowa, lecz jako funkcja wypukła i ograniczona przez pewną wartość. Praktycznie nie ma większego znaczenia, czy pewien potencjalny klient (lub dostawca) jest oddalony o, powiedzmy, tysiąc kilometrów czy o 10 tysięcy kilometrów. I tak jest nieosiągalny. Natomiast przyjęcie w funkcji (1) odległości Minkowskiego oznaczałoby, że klient oddalony o 10 tysięcy kilometrów powoduje 10 razy większe straty niż ten w odległości tysiąca – co jest oczywiście, z praktycznego punktu widzenia, bezzasadne.

Funkcję (1) przy odległości określonej wzorem (3) można traktować jako swoją miarę zagęszczenia ludności (zob. rys. 3). Ekstremum takiej funkcji położone jest w obszarze o stosunkowo największej liczbie ludności. Co charakterystyczne, ekstremów tych jest wiele. Znalezienie minimum globalnego wymaga więc porównania wartości funkcji (1) w wielu punktach. Może się też zdarzyć sytuacja, w której wartość najmniejszą funkcja osiąga w kilku różnych punktach jednocześnie. Zatem tego typu reprezentanci zbioru mają zdecydowanie odmienny charakter od

reprezentantów, takich jak mediana czy średnia arytmetyczna, a także zupełnie odmienne zastosowania. To wszystko uzasadnia przyjęcie następującej definicji, odróżniającej klasy tych reprezentantów.

Definicja 2. Reprezentant zbioru (X, w) jest **reprezentantem globalnym**, jeśli funkcja (1) jest funkcją wypukłą. Reprezentant jest **reprezentantem lokalnym**, jeśli funkcja (1) nie jest funkcją wypukłą.



Rys. 3. Warstwie funkcji (1) z odległością określoną wzorem (3)

Źródło: opracowanie własne.

Na zakończenie warto wspomnieć, że odległość zdefiniowana za pomocą wzoru (3) zależy od dodatniego parametru r . Przy małych wartościach tego parametru w przypadku zbioru gmin Polski reprezentantem będzie modalna, czyli lokalizacja miasta Warszawa. Wzrost wartości tego parametru powoduje, że reprezentantem staje się lokalizacja miasta Katowice. Oznacza to, że przy pewnych założeniach „środkiem Polski” mogą być Katowice(!). Istnieje prosta interpretacja tego faktu. Największą polską miejscowością jest Warszawa, ale największą aglomeracją jest właśnie aglomeracja katowicka.

Literatura

- Billingsley P., *Prawdopodobieństwo i miara*, PWN, Warszawa 1987.
Brandt S., *Analiza danych. Metody statystyczne i obliczeniowe*, PWN, Warszawa 1998.
Jajuga K., *Some Additions to the Problem of L_p – Norm Based Parameters*, [w:] Taksonomia 6, Prace Naukowe Akademii Ekonomicznej nr 817, AE, Wrocław 1999.
Maciuk A., *Bistandard*, [w:] *Ekonomia Matematyczna* 3, AE, Wrocław 1999.
Smoluk A., *Uwagi o modzie*, [w:] Taksonomia 4, AE, Wrocław 1997.

WHERE IS THE MIDDLE OF POLAND OR HOW TO CHOOSE THE REPRESENTATIVE OF THE SET

Summary

In many cases a problem appears how to choose a point which would represent a set in the best way. Depending on the context such a point is called middle of the set, parameter of location, estimator, best representative or simply a representative of a set. There are many ways of choosing such a representative. What is essential is that the choice as such is preceded by setting the criteria of choice which depend among others on the type of the issue. A representative commonly used in probability issues should not be automatically used in determinism issues – as finding optimal location.

In the article various representatives are presented as well as techniques and determinants of their choice. Specifically, local representatives are discussed – such that are dependent more on measure than size of the analyzed set. These issues are illustrated by the set of Polish communes. Where is the geographical middle of Poland? With which criteria can the city of Warsaw be the middle? Can they be set so that e.g. the city of Katowice is one? These are some interesting questions which can be answered analyzing the issues of choosing the representatives of the set.