DOI: 10.5604/01.3001.0014.1150

PRZEGLĄD STATYSTYCZNY
R. LXIII – ZESZYT 1 – 2016

EWA SZLACHTOWSKA[1], DANIEL KOSIOROWSKI[2], DOMINIK MIELCZAREK[3]

OCENA JAKOŚCI APLIKACYJNEJ ODPORNEGO ALGORYTMU ANALIZY SKUPIEŃ TCLUST NA PRZYKŁADZIE ZBIORU DANYCH DOTYCZĄCYCH JAKOŚCI POWIETRZA W KRAKOWIE[4] [5]

1. WPROWADZENIE

Metody wielowymiarowej analizy skupisk należą do procedur statystycznych najczęściej wykorzystywanych w praktyce gospodarczej. Mamy tutaj na uwadze m.in. eksploracyjną analizę danych – poszukiwanie modelu generującego dane ekonomiczne, wyodrębnianie typów klientów centrum handlowego bądź ostatnio, w kontekście badań prowadzonych w oparciu o tzw. wielkie zbiory danych – redukcję wymiaru zagadnienia statystycznego za pomocą tzw. mikroskupisk i danych funkcjonalnych (por. np. Jajuga, 1993; Krzyśko i inni, 2008; Walesiak, Gatnar, 2009; Kosiorowski i inni, 2015). Analiza skupień pozwala ogarnąć wielkie ilości danych będących w dyspozycji ekonomisty i spojrzeć na dane z właściwej perspektywy. Warto zwrócić uwagę, że jakość grupowania z wykorzystaniem tzw. klasycznych metod analizy skupień w krytycznym stopniu zależy od spełnienia założeń leżących u podłoża metod statystycznych stosowanych w ich obrębie (np. gdy zakładamy, że dane generuje mieszanina rozkładów normalnych). W praktyce badań ekonomicznych bardzo często mamy do czynienia z odstępstwem od przyjmowanych założeń.

Podejście odporne w modelowaniu statystycznym i analizie danych ma na celu zaproponowanie procedur statystycznych dających wiarygodne oszacowania, stanowiących użyteczne testy nie tylko w sytuacji, gdy dane generowane są przez zakładany

[1] Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie, Wydział Matematyki Stosowanej, Katedra Analizy Matematycznej, Matematyki Obliczeniowej i Metod Probabilistycznych, al. Mickiewicza 30, 30-059 Kraków, Polska, autor prowadzący korespondencję – e-mail: szlachto@agh.edu.pl.

[2] Uniwersytet Ekonomiczny w Krakowie, Wydział Zarządzania, Katedra Statystyki, ul. Rakowicka 27, 31-510 Kraków, Polska.

[3] Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie, Wydział Matematyki Stosowanej, Katedra Analizy Matematycznej, Matematyki Obliczeniowej i Metod Probabilistycznych, al. Mickiewicza 30, 30-059 Kraków, Polska.

[4] Autorzy uprzejmie dziękują za wsparcie finansowe ze strony UEK w Krakowie oraz AGH w Krakowie w postaci środków na utrzymanie potencjału badawczego 2015 i 2016.

[5] Niniejszy artykuł bezpośrednio nawiązuje do przewodu doktorskiego Ewy Szlachtowskiej pt. *Odporna analiza skupisk w badaniach nowej gospodarki*, otwartego na Uniwersytecie Ekonomicznym w Krakowie, Wydział Zarządzania, Katedra Statystyki.

przez procedurę statystyczną rozkład, ale także w sytuacji, gdy rozkład generujący dane nieco odbiega od zakładanego rozkładu (znajduje się w pewnym sąsiedztwie zakładanego modelu). Proponowana procedura statystyczna powinna posiadać dobre własności zarówno, gdy w próbie nie ma elementów odbiegających od zasadniczej części chmury danych (tzw. obserwacji odstających, ang. *outliers*), ale także w sytuacji, gdy takie elementy występują (por. Maronna i inni, 2006). W ostatnich latach w literaturze proponuje się także odporne metody analizy skupień (por. np. Rocke, Wodurf, 2002; Kosiorowski, 2008). W związku z szeregiem trudności koncepcyjnych związanych z samym rozumieniem odporności metody analizy skupień – tematyka obfituje w szereg nierozwiązanych jak dotąd otwartych problemów (np. jak rozumieć, że procedura łamie się, w oparciu o którą miarę jakości grupowania danych definiować miarę wpływu jednostek odstających na wynik grupowania, czy odporność procedury wiązać z liczbą bądź kształtem skupień powstających w wyniku jej działania itd.)

W niniejszym artykule skupiono się na niehierarchicznym odpornym algorytmie grupowania o nazwie TCLUST, który jest jednym z najlepszych zaproponowanych jak dotąd w literaturze. Jego odporność rozumiemy w ramach jednolitego podejścia zaproponowanego w Genton, Lucas (2003).

2. ALGORYTM TCLUST

Odporny algorytm TCLUST mający swoją darmową implementację w postaci pakietu środowiska R (R Development Core Team 2010) o tej samej nazwie dobrze radzi sobie z niedoskonałościami danych ekonomicznych. Twórcami algorytmu TCLUST są H. Fritz, L. A. García-Escudero, A. Mayo-Iscar (por. Fritz i inni, 2011; Fritz i inni, 2012). Pakiet tclust dostępny jest na stronie http://CRAN.R-project.org/package=tclust.

Założenia modelu są następujące. Rozważamy próbę losową $\{x_1,\ldots,x_n\} \subset \mathbb{R}^p$, gdzie x_i ($i = 1,\ldots,n$) są zmiennymi losowymi o p-wymiarowym rozkładzie normalnym z funkcją gęstości $\phi(\cdot,\mu,\Sigma)$, wektorem wartości oczekiwanych $\mu = [\mu_1,\ldots,\mu_p]^T$, i macierzą kowariancji Σ. Szukamy podziału $\{R_0,R_1,\ldots,R_k\}$ zbioru indeksów $\{1,\ldots,n\}$, gdzie $\#R_0 = \lceil n\alpha \rceil$, środków (centroidów) $m_1,\ldots,m_k$, symetrycznych dodatnich półokreślonych macierzy rozproszenia $S_1,\ldots,S_k$ oraz wag $p_1,\ldots,p_k$, gdzie $p_j \in [0,1]$ i $\sum_{j-1}^{k} p_j = 1$ maksymalizujących funkcję celu:

$$\sum_{j=1}^{k} \sum_{i \in R_j} \log\left(p_j \phi(x_i, m_j, S_j) \right). \tag{1}$$

Rozważamy „niejednorodny" problem grupowania obserwacji, tzn. dopuszczamy skupienia o eliptycznym kształcie oraz dopuszczamy istnienie małej α części obserwacji, które mogłyby wpłynąć negatywnie na jakość podziału danych na skupienia.

Algorytm TCLUST łączy w sobie możliwości metody grupowania k-średnich z możliwością odpornego oszacowania macierzy kowariancji, jaką daje algorytm wyznaczania estymatora minimalnego wyznacznika macierzy kowariancji (ang. *fast--MCD algorithm*, por. Rousseeuw, Van Driessen, 1999). Nowe środki oraz nowe macierze rozproszenia są wyznaczane poprzez wyliczenie empirycznego wektora średnich i empirycznej macierzy kowariancji obserwacji przypisanych do każdego skupienia. Niestety, takie połączenie obu algorytmów nie zapewnia rozsądnych wyników podziału danych. Spowodowane jest to tym, że duże skupienia mają tendencję do „pochłaniania" najmniejszych. Aby wyeliminować działanie niepożądane algorytmu analizy skupień, nakłada się pewne ograniczenia kontrolujące rozproszenie w skupieniu. Zauważmy również, że problem bezpośredniej maksymalizacji funkcji danej wzorem (1), bez nałożenia ograniczeń na macierze rozproszenia, nie jest dobrze określony. Przykładowo, gdy dla macierzy rozproszenia S_j zachodzi $\det(S_j) \rightarrow 0$, wówczas funkcja celu jest nieograniczona. W związku z tym, aby problem maksymalizacji funkcji celu danej wzorem (1) był dobrze określony, w pracy García-Escudero i inni (2008) został nałożony warunek na iloraz wartości własnych macierzy rozproszenia:

$$\frac{\max_{j,l} \lambda_{j,l}}{\min_{j,l} \lambda_{j,l}} \leq c, \tag{2}$$

gdzie $\lambda_{j,l}$ są wartościami własnymi ($l = 1,\ldots,p$) macierzy rozproszenia S_j ($j = 1,\ldots,k$), a $c \geq 1$ jest stałą, która kontroluje siłę ograniczenia (2). Zauważmy, że im większa zostanie wybrana stała c, tym „luźniejsze" jest ograniczenie macierzy rozproszenia, co pozwala na większą różnorodność wśród skupień. Przeciwnie, dzięki małej wartości stałej c (blisko jeden) otrzymamy bardziej jednorodnie „rozproszone" skupienia. Dla $c = 1$, otrzymujemy ważoną metodę przyciętych k-średnich oraz skupienia o kulistym kształcie.

Algorytm TCLUST, maksymalizujący funkcję celu (1) przy warunku (2), opiera się głównie na stosowaniu pewnych kroków szacowania (E-krok, ang. *estimation*) i maksymalizacji (M-krok, ang. *maximization*). Algorytm można podzielić na dwa etapy: inicjalizacji oraz zagęszczania (ang. *concentration step*). Na etapie inicjalizacji losowo wybieramy k początkowych środków m_j^0, k początkowych macierzy rozproszenia S_j^0 oraz k początkowych wag p_j^0. W rezultacie otrzymujemy wektor $\theta^0 = (p_1^0,\ldots,p_k^0,m_1^0,\ldots,m_k^0,\ldots,S_1^0,\ldots,S_k^0)$. Jeśli macierze S_j^0 nie spełniają ograniczenia na iloraz wartości własnych, możemy po prostu wziąć wszystkie S_j^0 równe macierzy jednostkowej i/lub zmodyfikować otrzymane macierze rozproszenia tak, aby spełniały ograniczenia nałożone na iloraz wartości własnych.

Etap zagęszczania składa się z przypisywania obserwacji do skupień, a następnie aktualizacji parametrów. Następujące kroki wykonujemy naprzemiennie, aż nie nastąpią żadne zmiany (czyli $\theta^l = \theta^{l+1}$) lub jeżeli zostanie wykonana z góry zadana liczba kroków. W E-kroku, w danej iteracji przypisanie do skupień odbywa się poprzez pomiar odległości obserwacji x_i od każdego środka m_j. Ponieważ dopuszczamy różne

wagi i macierze rozproszenia, funkcje mierzące odległość od środków definiujemy jako:

$$D_j(x_i,\theta^l) = p_j^l \phi(x_i, m_j^l, S_j^l), \tag{3}$$

gdzie $\theta^l = (p_1^l,\ldots,p_k^l,m_1^l,\ldots,m_k^l,\ldots,S_1^l,\ldots,S_k^l)$ to zbiór parametrów w bieżącej iteracji algorytmu, tj. zbiór środków $m_1^l,\ldots,m_k^l$, symetrycznych dodatnich półokreślonych macierzy rozproszenia $S_1^l,\ldots,S_k^l$ oraz wag $p_1^l,\ldots,p_k^l$, gdzie $p_k^l \in [0,1]$ i $\Sigma\, p_j^l = 1$. Funkcje D_j nazywamy funkcjami dyskryminującymi (ang. *discriminant functions*). Najmniejsze $D_j(x_i,\theta)$ oznacza, że obserwacja x_i jest najbardziej odległa od środka m_j. Ponadto będziemy rozważać miarę „odstawania obserwacji" (ang. *outlyingness measure*) jako $D(x_i,\theta) = \max\{D_1(x_i,\theta^l),\ldots, D_k(x_i,\theta^l)\}$. Na podstawie wartości miary odstawania dla poszczególnych obserwacji, wybieramy zbiór H składający się z indeksów 1–α części obserwacji o największych wartościach $D(x_i,\theta^l)$. Następnie dzielimy H na podzbiory $H_1,\ldots,H_k$, gdzie $H_j = \{i \in H : D_j(x_i,\theta^l) = D(x_i,\theta^l)\}$. Może się zdarzyć, że otrzymamy puste skupienie albo obserwację, która może należeć do więcej niż jednego skupienia. Wtedy przyjmujemy zasadę, że taką obserwację przypisujemy do skupienia o najniższym indeksie. Warto podkreślić, że dla k = 1 odległość dana wzorem (3) jest odległością Mahalanobisa wykorzystywaną w algorytmie fast-MCD. Natomiast dla $p_1 = \ldots = p_k$ oraz $S_1 = \ldots S_k = \sigma^2 I$ jest to odległość stosowana w metodzie k-średnich.

W M-kroku aktualizujemy parametry korzystając z informacji przypisania obserwacji do skupień. W tym kroku bardzo ważne jest nałożenie ograniczenia na iloraz wartości własnych macierzy rozproszenia i jego kontrola. Sposób nakładania ograniczeń jest jednym z najważniejszych etapów w algorytmie. Jednakże jest to „wąskie gardło" obliczeniowe algorytmu, ponieważ złożony problem optymalizacji musi być rozwiązywany w każdej iteracji zagęszczania. Aktualizacja parametrów przebiega w następujący sposób. Wagi są aktualizowane przez $p_j^{l+1} = n_j / [n(1-\alpha)]$, gdzie $n_j = \#H_j$. Natomiast środki dane są jako empiryczne wektory średnich

$$m_j^{l+1} = \frac{1}{n_j} \sum_{i \in H_j} x_i. \tag{4}$$

Aktualizacja estymatorów rozproszenia nie jest taka prosta, ponieważ przy obliczaniu estymatorów macierzy kowariancji

$$T_j = \frac{1}{n_j} \sum_{i \in H_j} \left(x_i - m_j^{l+1}\right)\left(x_i - m_j^{l+1}\right)^T \tag{5}$$

może się zdarzyć, że macierze $T_1,\ldots,T_k$ nie spełniają ograniczenia (2) nałożonego na iloraz wartości własnych. W takim przypadku, rozważamy rozkład macierzy $T_j = U_j^T D_j U_j$, gdzie U_j jest ortogonalną macierzą, a $D_j = \mathrm{diag}(d_{j1},d_{j2},\ldots,d_{jp})$ macierzą diagonalną. Zdefiniujmy odcięte wartości własne jako

$$d_{jl}^{m} = \begin{cases} d_{jl} & \text{dla } d_{jl} \in [m, cm] \\ m & \text{dla } d_{jl} < m \\ cm & \text{dla } d_{jl} > cm \end{cases}, \tag{6}$$

gdzie m > 0. Macierze rozproszenia aktualizujemy w następujący sposób: $S_j^{l+1} = U_j^T D_j^* U_j$, gdzie $D_j^* = \text{diag}(d_{j1}^{m_{opt}}, d_{j2}^{m_{opt}}, \ldots, d_{jp}^{m_{opt}})$, a m_{opt} jest wartością m (m > 0) minimalizującą wyrażenie

$$m \mapsto \sum_{j=1}^{k} n_j \sum_{l=1}^{p} \left(\log(d_{jl}) + \frac{d_{ij}^{m}}{d_{ij}} \right). \tag{7}$$

Zauważmy, że w każdym kroku zagęszczania następuje wzrost wartości funkcji celu. Po przeprowadzeniu wszystkich kroków algorytmu wartość funkcji celu (1) jest wyliczana. Natomiast wynikiem końcowym algorytmu jest zbiór parametrów, które prowadzą do jak największej wartości funkcji celu.

3. WYBÓR LICZBY PARAMETRÓW POCZĄTKOWYCH

Prawdopodobnie jednym z najbardziej złożonych problemów przy stosowaniu analizy skupień jest wybór liczby skupień. Wybór liczby skupień k oraz wybór poziomu przycinania α są to powiązane problemy, które powinny być rozwiązywane jednocześnie. Ważne jest, aby zauważyć, że dany poziom przycinania pociąga za sobą określoną liczbę skupień i vice versa. Ta zależność związana jest między innymi z tym, że całe skupienia mogą zostać całkowicie przycięte przy zbytnim zwiększeniu α. Z drugiej strony, gdy wybrany poziom α jest zbyt niski, to grupy obserwacji odstających mogą tworzyć nowe fałszywe skupienia. W rezultacie wydaje się, że liczba skupień występująca w zbiorze danych jest wyższa. Ponadto równoczesny wybór k i α zależy od rodzaju skupień, których szukamy jak i dopuszczalnych różnic pomiędzy rozmiarami skupień. W pracy García-Escudero i inni (2011) zostało zaproponowane pewne narzędzie graficzne pomagające dokonać właściwego wyboru liczby skupień k oraz poziomu przycinania α.

Załóżmy najpierw, że stała c została wcześniej ustalona przez badacza, który stosuje metody odpornej analizy skupień. Tradycyjną metodą wyboru liczby skupień, gdy α = 0, jest uważne kontrolowanie wartości maksymalnej funkcji celu. Jednakże zwiększanie liczby skupień k zawsze spowoduje wzrost maksymalnej wartości funkcji (1), co może prowadzić do „przeszacowania" liczby skupień. Do podejmowania rozsądnych wyborów parametrów α i k, w pracy Fritz i inni (2012) proponuje się monitorowanie funkcji największej wiarygodności $(\alpha,k) \rightarrow \mathcal{L}(\alpha,k)$, gdzie $\mathcal{L}(\alpha,k)$ jest maksymalną wartość osiągniętą przez funkcję (1) dla każdej kombinacji z danego

zbioru wartości dla k i α. W praktyce proponuje się, aby wybrać liczbę skupień jako najmniejszą wartość k taką, że

$$\mathcal{L}(\alpha,k+1) - \mathcal{L}(\alpha,k) \tag{8}$$

jest (prawie) 0 z wyjątkiem niewielkich wartości α. Gdy liczba skupień jest ustalona, jako poziom przycinania wybieramy pierwsze α_0 takie, że (8) jest bliskie 0 dla każdego $\alpha \geq \alpha_0$.

W pakiecie tclust jest dostępna funkcja ctlcurves, która przybliża funkcję największej wiarygodności poprzez wykonanie funkcji tclust dla sekwencji wartości k i α. Opisana powyżej procedura, dokonywania rozsądnych wyborów parametrów k i α, wymaga aktywnego udziału badacza. Wielkość ograniczenia na iloraz wartości własnych musi być z góry określona. W konsekwencji decyzja badacza, dotycząca wielkości nałożonych ograniczeń, determinuje właściwe ustalenie parametrów k i α. Na przykład, niektóre specyficzne zastosowania analizy skupień wymagają niemal kulistych skupień, które mogą być uzyskane przez ustalenie stałej c blisko 1. Domyślnie jest ustawiona wartość c = 50, jednakże można w razie potrzeby zmienić wartość stałej c.

Otrzymane w ten sposób „rozsądne" wartości k i α oraz związane z nimi skupienia należy dokładnie zbadać. Ponadto algorytm TCLUST wysyła ostrzeżenie, gdy uzyskane skupienia zostały „sztucznie ograniczone" algorytmem. Oznacza to, że wartości własne macierzy rozproszenia spełniają warunek

$$\frac{\max\limits_{j,l} \lambda_{j,l}}{\min\limits_{j,l} \lambda_{j,l}} = c, \tag{9}$$

ponieważ algorytm wymusił takie ograniczenie. W tej sytuacji, jeśli nie są wymagane żadne szczególne ograniczenia, stała c może być stopniowo zwiększana do czasu, aż ostrzeżenie zniknie.

Ponadto w pakiecie tclust dostępne jest dodatkowe narzędzie graficzne, które może być stosowane w celu oceny jakości podziału danych i podjętych decyzji przycinania. Oceny tej dokonuje się przez zastosowanie funkcji dyskryminujących.

Niech $R = \{R_0,R_1,\ldots,R_k\}$, $\theta = (p_1,\ldots,p_k,m_1,\ldots,m_k,\ldots,S_1,\ldots,S_k)$ będą wartościami otrzymanymi z algorytmu poprzez maksymalizację funkcji (1). $D_j(x_i,\theta)$ jest miarą stopnia przynależności obserwacji x_i do j-tego skupienia. Wartości funkcji dyskryminujących, mierzących odległości obserwacji x_i od poszczególnych środków skupień, można posortować rosnąco

$$D_{(1)}(x_i,\theta) \leq \ldots \leq D_{(k)}(x_i,\theta), \tag{10}$$

gdzie $D_{(k)}(x_i,\theta) = D(x_i,\theta)$. Jakość decyzji przyporządkowania nieprzyciętych obserwacji x_i można ocenić poprzez porównanie jej stopnia przynależności do najbliższego skupienia z drugim z możliwych najlepszych dopasowań, tj.

$$DF(i) = \log\left(\frac{D_{(k-1)}(x_i,\theta)}{D_{(k)}(x_i,\theta)}\right). \tag{11}$$

Ponadto jeżeli $x_{(1)},\ldots,x_{(n)}$ będą obserwacjami w próbie posortowanymi względem ich wartości $D_{(k)}(\cdot,\theta)$ ($D_{(k)}(x_{(1)},\theta) \leq \ldots \leq D_{(k)}(x_{(n)},\theta)$) to $x_{(1)},\ldots,x_{(\lceil n\alpha\rceil)}$ są to przycięte obserwacje, które nie są przypisane do żadnego skupienia. Niemniej jednak, możliwe jest obliczenie stopnia przynależności przyciętej obserwacji x_i do jej najbliższego skupienia. Tym samym jakość decyzji przycinania owej obserwacji można ocenić przez porównanie $D_{(k)}(x_i,\theta)$ z $D_{(k)}(x_{\lceil n\alpha\rceil+1},\theta)$, tj.

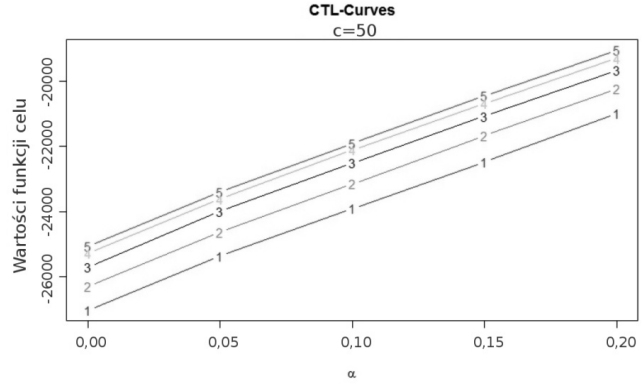
$$DF(i) = \log\left(\frac{D_{(k)}(x_i,\theta)}{D_{(k)}(x_{\lceil n\alpha+1\rceil},\theta)}\right). \tag{12}$$

W rezultacie czynniki dyskryminujące $\mathrm{DF}(i) \leq 0$ są uzyskiwane dla każdej obserwacji w zbiorze danych, przyciętej czy też nie. Do oceny uzyskanych czynników dyskryminujących służy funkcja DiscrFact. Wynikiem funkcji DiscrFact są trzy wykresy. Wykres „Klasyfikacja" (ang. *classification*) ilustruje przypisanie obserwacji do skupień oraz podjęte decyzje przycinania. Wykres zarysu (ang. *silhouette plot*, por. Rousseeuw, 1987) przedstawia wartości czynników dyskryminujących dla poszczególnych skupień. Czynniki dyskryminujące określają dobroć podjętej decyzji grupowania. Obiekty na wykresie zarysu z wieloma dużymi wartościami $\mathrm{DF}(i)$ (tj. blisko 0) wskazują na istnienie niezbyt „dobrze dobranych" skupień. Najbardziej „wątpliwe" przypisania o czynniku $\mathrm{DF}(i)$ większym niż ustalona wartość progowa ($\log(prog)$) są wyświetlane przez funkcję DiscrFact. Przykładowo, wybór $prog = 0{,}1$ oznacza, że decyzja dla danej obserwacji x_i jest uważana za wątpliwą, jeżeli wielkość czynnika dyskryminującego drugiej najlepszej z możliwych decyzji ($D_{(k-1)}(x_i,\theta)$ lub $D_{(k)}(x_{\lceil n\alpha\rceil+1},\theta)$) jest większa niż jedna dziesiąta wielkości czynnika dyskryminującego rzeczywiście podjętej decyzji ($D_{(k)}(x_i,\theta)$). Natomiast im mniejsze wartości czynników dyskryminujących tym lepiej dopasowane skupienia. Wątpliwe decyzje są zaznaczone na wykresie „Obserwacje o wątpliwej przynależności" (ang. *doubtful assignments*).

4. ZASTOSOWANIE ALGORYTMU TCLUST DLA ZBIORU DANYCH DOTYCZĄCYCH JAKOŚCI POWIETRZA

Aby zobrazować wady i zalety algorytmu TCLUST wybrano dane dotyczące jakości powietrza ze stacji Aleja Krasińskiego w Krakowie za okres od 1 do 31 października 2015 r. Wyniki pomiarów siedmiu substancji są prezentowane co godzinę na stronie http://monitoring.krakow.pios.gov.pl/.

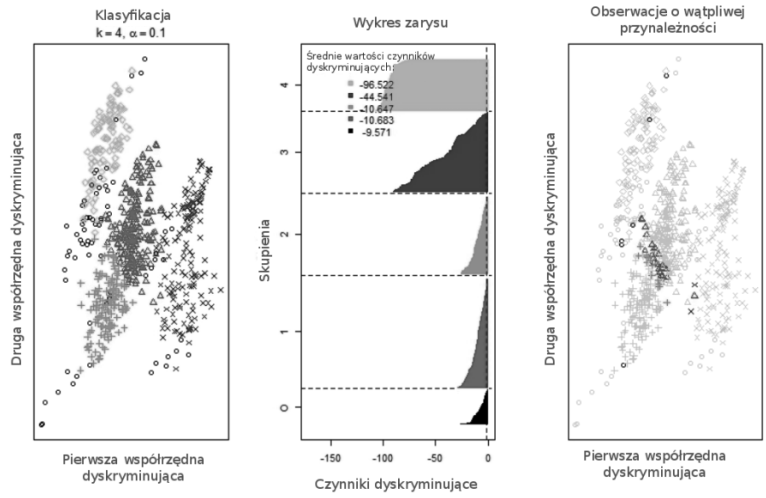
W związku z próbą pogrupowania otrzymanych danych ze względu na wielkość poziomu niektórych substancji w powietrzu w zależności od pory dnia zastosowano algorytm TCLUST. W celu ustalenia optymalnej liczby skupień zastosowano funkcję ctlcurves dostępną w pakiecie tclust dla sekwencji parametrów: k od 1 do 5 oraz α od 0 do 0,2.



Rysunek 1. Klasyfikacja funkcji największej wiarygodności dla parametrów k = 1,…,5, α = 0, 0,05,…, 0,2

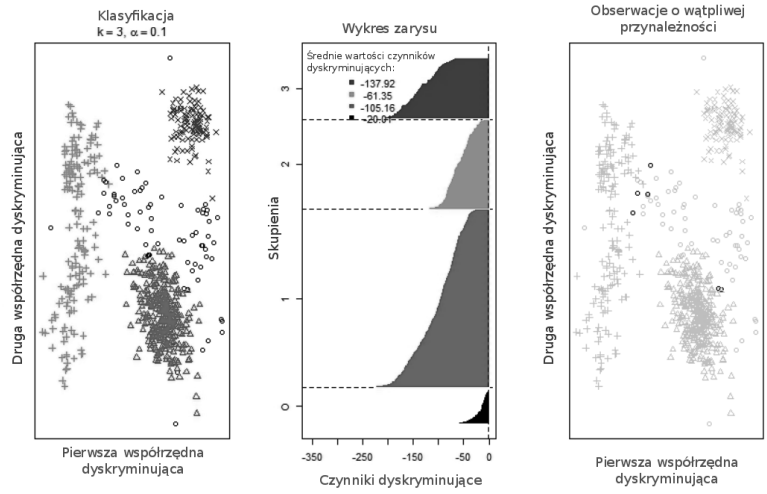
Źródło: opracowanie własne za pomocą programu R.

Na rysunku 1 widać wyraźną różnicę pomiędzy wartościami funkcji celu obliczonymi dla jednego skupienia oraz dwóch skupień. Również widoczny wzrost jest wartości funkcji celu przechodząc od dwóch skupień do trzech, jak również od trzech do czterech. Jednakże trudno określić czy rozważać cztery skupienia, czy konieczne jest rozważanie pięciu skupień. Podobnie, różnica w wartościach funkcji celu dla różnych poziomów przycinania, przechodząc od trzech do czterech skupień wydaje się praktycznie stała dla uwzględnionych wartości przycinania.



Rysunek 2. Wykres zarysu dla parametrów k = 4, α = 0,1, c = 50
Źródło: opracowanie własne za pomocą programu R.

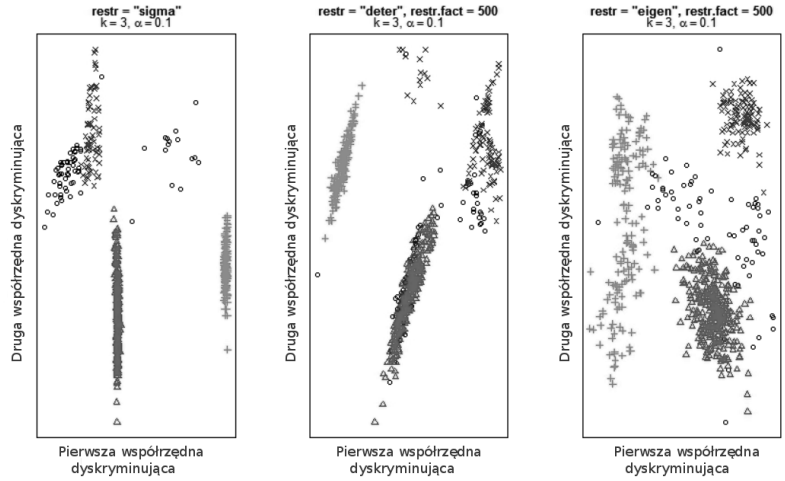
Rysunek 2 przedstawia wykres zarysu. Na wykresie „Klasyfikacja" można zaobserwować, że wątpliwe obserwacje występują głównie na styku dwóch skupień. Wykres zarysu przedstawia wartości czynników dyskryminujących dla poszczególnych skupień. Dla rozważanych parametrów k = 4, α = 0,1 oraz c = 50 wartości czynników dyskryminujących tylko w dwóch skupieniach są oddalone od zera, co świadczy o nienajlepszym wyborze parametrów. Dodatkowo otrzymaliśmy informację, iż algorytm wymusił ograniczenie na iloraz wartości własnych macierzy rozproszenia.



Rysunek 3. Wykres zarysu dla parametrów k = 3, α = 0,1, c = 500
Źródło: opracowanie własne za pomocą programu R.

Na rysunku 3 został przedstawiony wykres zarysu dla parametrów k = 3, α = 0,1, c = 500. Na wykresie „Obserwacje o wątpliwej przynależności" niewiele decyzji zostało zaznaczonych jako decyzje wątpliwe. Natomiast na wykresie zarysu wartości czynników dyskryminujących są oddalone od zera, co świadczy o dobrym doborze parametrów. Warto podkreślić, że mimo sugestii, iż dla zestawu parametrów k = 4 i α = 0,1 wartość funkcji celu będzie znacząco wyższa, to jednak zestaw parametrów k = 3, α = 0,1 wskazuje na lepsze decyzje grupowania.

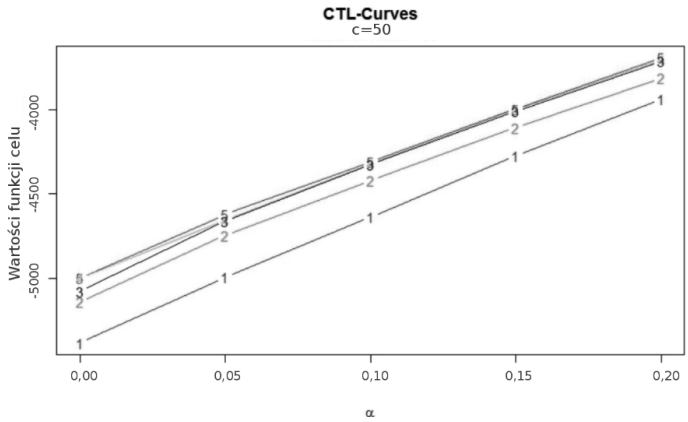
W pakiecie tclust istnieje również możliwość nałożenia innych ograniczeń na macierze rozproszenia. Oprócz ograniczenia na iloraz wartości własnych, które to decyduje o kształcie skupień, można zażądać, aby wszystkie macierze rozproszenia były takie same albo nałożyć ograniczenie na wyznaczniki macierzy rozproszenia. Tego typu ograniczenia wpływają na objętość skupień, a nie na ich kształt.



Rysunek 4. Wyniki grupowania przy różnych ograniczeniach na macierze rozproszenia dla parametrów k = 3, α = 0,1, c = 500

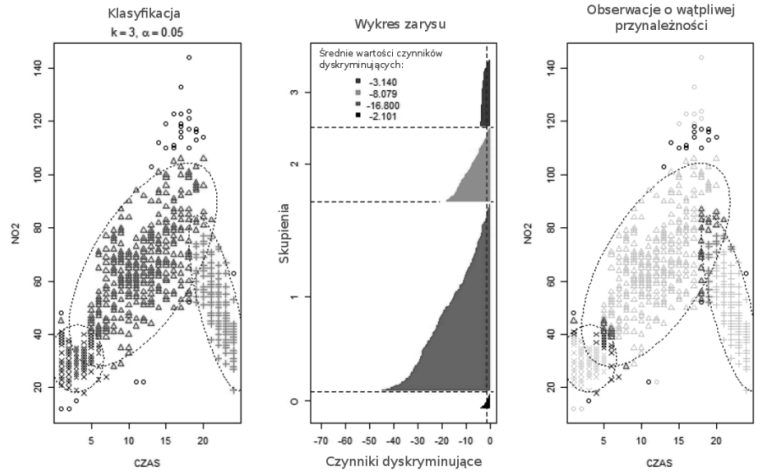
Źródło: opracowanie własne za pomocą programu R.

Po wstępnej analizie zbioru danych dotyczących jakości powietrza, trudno znaleźć zależność pomiędzy porą dnia a przypisaniem do skupień. Otrzymany podział wymaga dalszej i bardziej szczegółowej analizy dotyczącej jakości powietrza. W związku z tym w kolejnym etapie ograniczyliśmy się tylko do dwutlenku azotu.



Rysunek 5. Klasyfikacja funkcji największej wiarygodności dla parametrów k = 1,…,5, α = 0, 0,05,…, 0,2
Źródło: opracowanie własne za pomocą programu R.

Na rysunku 5 widać wyraźny wzrost wartości funkcji celu przechodząc od jednego skupienia do dwóch, jak również od dwóch do trzech. W związku z tym zastosowano algorytm TCLUST dla parametrów k = 3 i α = 0,05.



Rysunek 6. Wykres zarysu dla parametrów k = 3, α = 0,05, c = 500
Źródło: opracowanie własne za pomocą programu R.

Rysunek 6 przedstawia wykres zarysu. Na wykresie „Klasyfikacja" można zaobserwować, że wątpliwe obserwacje występują głównie na styku skupień. Ponadto na rysunku 6 można zaobserwować, że między godziną 6 a 19 występuje największe natężenie emisji dwutlenku azotu. Na szczególną uwagę zasługują również obserwacje odstające, które powinny zostać poddane dalszej szczegółowej analizie. W tabeli 1

podano środki poszczególnych skupień. Można zaobserwować, że koło południa występuje najwyższe stężenie dwutlenku azotu i utrzymuje się do późnych godzin wieczornych. Natomiast tuż nad ranem jest najniższe.

Tabela 1.

Środki poszczególnych skupień dla parametrów k = 3, α = 0,05, c = 500

Środki skupień	C1	C2	C3
Czas	11,91	21,84	3,04
NO_2	66,17	51,77	30,99

Źródło: opracowanie własne za pomocą programu R.

5. PODSUMOWANIE

Jak już wcześniej wspomniano, w analizie skupień bardzo istotna jest rola badacza. W pierwszym kroku należy wybrać odpowiedni algorytm analizy skupień. Tutaj istotne jest jaką informację o zbiorze danych posiada badacz. Pakiet tclust służy do analizy bardzo dużych zbiorów danych o różnej wymiarowości. Algorytm TCLUST prezentuje podejście odporne w modelowaniu statystycznym. Ponadto zakłada się między innymi, że dane generowane są przez rozkład normalny lub zbiór danych jest mieszaniną rozkładów normalnych. W związku z tym badacz powinien mieć pewną informację na wejściu dotyczącą danych. Bardzo dobrze sprawdza się w sytuacji, gdy szukamy niejednorodnych skupień o eliptycznych kształtach. Dzięki możliwości nakładania pewnych ograniczeń na macierze rozproszenia, można wpływać na kształt, jak również na objętość otrzymanych skupień. Zauważmy, że $\lceil n\alpha \rceil$ obserwacji nie jest brane pod uwagę przy obliczaniu funkcji celu (1). W rezultacie algorytm dobrze się sprawdza w sytuacji, gdy rozkład generujący dane nieco odbiega od zakładanego rozkładu, jak również można uniknąć szkodliwego wpływu odstających obserwacji. Kolejnym krokiem jest dobór parametrów wejściowych: liczba skupień, poziom przycinania, wartość ograniczenia na iloraz wartości własnych. W pakiecie tclust zaproponowano kilka rozwiązań pomocnych przy wyborze parametrów wejściowych. Jest to funkcja ctlcurves, która służy do wyboru liczby skupień oraz poziomu przycinania, oraz funkcja DicsrFact, która służy do weryfikacji podjętych decyzji grupowania. Ponadto algorytm informuje badacza czy nałożone ograniczenie na iloraz wartości własnych macierzy rozproszenia zostało „wymuszone". Ostatnim i najtrudniejszym etapem grupowania jest interpretacja wyników. Przykładowo, po wstępnej analizie zbioru danych dotyczących jakości powietrza, trudno znaleźć zależność pomiędzy porą dnia a przypisaniem do skupień. Otrzymany podział wymagał dalszej i bardziej szczegółowej analizy danego zbioru danych. Jednakże po ograniczeniu się do danych pomiarowych dotyczących zanieczyszczeń powietrza dwutlenkiem azotu udało się znaleźć zależność pomiędzy porą dnia a wielkością stężenia NO_2 w Krakowie.

LITERATURA

Fritz H., García-Escudero L. A., Mayo-Iscar A., (2011), A Fast Algorithm for Robust Constrained Clustering, URL http://www.eio.uva.es/infor/personas/tclust_algorithm.pdf.

Fritz H., García-Escudero L. A., Mayo-Iscar A., (2012), tclust: An R Package for a Trimming Approach to Cluster Analysis, *Journal of Statistical Software*, 47 (12), 1–26.

García-Escudero L. A., Gordaliza A., Matrán C., Mayo-Iscar A., (2011), Exploring the Number of Groups in Robust Model-Based Clustering, *Statistics and Computing*, 21 (4), 585–599.

García-Escudero L. A., Gordaliza A., Matrán C., Mayo-Iscar A., (2008), A General Trimming Approach to Robust Cluster Analysis, *The Annals of Statistics*, 36 (3), 1324–1345.

Genton M. G., Lucas A., (2003), Comprehensive Definitions of Breakdown Points for Independent and Dependent Observations, *Journal of the Royal Statistical Society Series B*, 65, 81–84.

Jajuga K., (1993), *Statystyczna analiza wielowymiarowa*, PWN, Warszawa.

Kosiorowski D., Mielczarek D., Szlachtowska E., (2015), Clustering of Functional Objects in Energy Load Prediction Issues, w: Papież M., Śmiech S., (red.), *Proceedings from 9th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena,* Fundacja Uniwersytetu Ekonomicznego w Krakowie, 108–118.

Kosiorowski D., Zawadzki Z., (2014), *DepthProc*: An R Package for Robust Exploration of Multidimensional Economic Phenomena, http://arxiv.org/abs/1408.4542.

Kosiorowski D., (2008), Robust Classification and Clustering Based on the Projection Depth Function, w: Brito P., (red.), COMPSTAT 2008, Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, 209–216.

Krzyśko M., Wołyński W., Górecki T., Skorzybut M., (2008), *Systemy uczące się*, WNT.

Maronna R. A., Martin R. D., Yohai V. J, (2006), *Robust Statistics – Theory and Methods,* John Wiley & Sons, Chichester.

Rocke D. M., Woodruf D. L., (2002), Computational Connections Between Robust Multivariate Analysis and Clustering, w: Härdle R. B., (red.), *COMPSTAT 2002 Proceedings in Computational Statistics*, 255–260.

Rousseeuw, P. J., Van Driessen K., (1999), A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics,* 41 (3), 212–223.

Rousseeuw P. J., (1987), Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, *Journal of Computational and Applied Mathematics*, 20 (1), 53–65.

Walesiak M., Gatnar E., (red.), (2009), *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa.

OCENA JAKOŚCI APLIKACYJNEJ ODPORNEGO ALGORYTMU ANALIZY SKUPIEŃ TCLUST NA PRZYKŁADZIE ZBIORU DANYCH DOTYCZĄCYCH JAKOŚCI POWIETRZA W KRAKOWIE

Streszczenie

Pozyskiwanie i gromadzenie danych to obecnie bardzo dynamiczne procesy. Przy ogromnych ilościach danych proces przetwarzania danych w celu uzyskania na ich podstawie użytecznych informacji i wniosków nie jest zadaniem trywialnym. W tym pomocna jest analiza skupień, a wynik grupowania pozwala ogarnąć dostępną informację i spojrzeć na nią z innej perspektywy. W żadnym razie nie jesteśmy w stanie pokazać całego spektrum zagadnień związanych analizą skupień, dlatego też ograniczymy się do omówienia algorytmu TCULST, którego twórcami są H. Fritz, L. A. García-Escudero, A. Mayo-Iscar (por. Fritz i in., 2011, 2012). W pracy zostaną przedstawione wady i zalety odpornego algorytmu

analizy skupień oraz omówione podstawowe funkcje dostępne w pakiecie tclust. Następnie zostanie dokonana ocena jakości aplikacyjnej algorytmu TCLUST na przykładzie zbioru danych dotyczących jakości powietrza w Krakowie.

Słowa kluczowe: odporna analiza skupień, algorytm tclust, badanie jakości powietrza

EVALUATION OF THE QUALITY OF ROBUST CLUSTERING ALGORITHM TCLUST ON THE EXAMPLE OF DATASET OF AIR POLLUTANTS EMISSION IN KRAKOW

Abstract

Acquisition and data collection is currently a very dynamic processes. In order to obtain from data useful information, when huge quantities of data, the processing of the data is not a trivial task. Cluster analysis is very helpful in this and the result of grouping the result of grouping allows us to comprehend the available information and look at it from a different perspective. In any case, we are not able to show the entire spectrum of issues related to data analysis. Therefore we limit our discussion to the analysis of clusters, then we describe the TCLUST algorithm. The authors of the algorithm are H. Fritz, L. A. García-Escudero, A. Mayo-Iscar (see Fritz et al. 2011, 2012). In the paper we present the pros and cons robust clustering algorithm, and we discuss the available functions in the package tclust. Then on the example of dataset of air pollutants emission in Krakow we try to evaluate the quality of robust clustering algorithm.

Keywords: robust cluster analysis, tclust algorithm, air quality testing