УДК 519.81 **DOI: 10.20998/2222-0631.2021.02.12**

М. С. СОФРОНОВА

МЕТОД УСУНЕННЯ АНОМАЛЬНИХ ВИМІРЮВАНЬ ПРИ АНАЛІЗІ БАЗИ БАГАТОВИМІРНИХ ДАНИХ ПІД ЧАС РОЗВ'ЯЗАННЯ ЗАДАЧІ ПРИЙНЯТТЯ РІШЕНЬ

У роботі запропоновано метод усунення аномальних вимірювань (викидів) для підвищення рівня якості багатовимірних даних при статистичних дослідженнях. Така проблема виникає, наприклад, в теорії прийняття управлінських рішень, оскільки при обчисленні оцінок параметрів імовірнісних розподілів наявність у вибірці аномальних (тобто таких, що значно збільшують довірчий інтервал) вимірювань здатна спотворити результати статистичного дослідження, а, отже, і основної задачі. Особливість запропонованого метода полягає в тому, що він є комбінацією статистичних та геометричного методів, а саме: методу оцінювання Гествірта, процедури Тьюкі та модифікації метода побудови опуклої оболонки скінченої множини точок багатовимірного простору. Множині багатовимірних даних ставиться у відповідність множина точок багатовимірного простору. Для знаходження і виключення викидів будується послідовність вкладених опуклих оболонок – n – політопів, кожен з яких описується перетином напівпросторів (опорних гіперграней). Наводиться детальний алгоритм знаходження аномальних вимірювань. Їх виключення відповідає послідовному виключенню граничних точок вкладених опуклих оболонок. Оцінка Гествірта дає умову зупинки роботи алгоритму. Запропонований метод не потребує великих обчислювальних витрат та може широко використовуватися при розв'язанні як теоретичних, так і практичних задач, пов'язаних з обробкою багатовимірних даних. Наведено чисельні результати роботи методу з кількістю компонент даних 4 та 5.

Ключові слова: статистичне дослідження, теорія прийняття рішень, аномальні вимірювання, викиди, опукла оболонка, n – вимірний політоп, багатовимірний простір.

М. С. СОФРОНОВА

МЕТОД УСТРАНЕНИЯ АНОМАЛЬНЫХ ИЗМЕРЕНИЙ ПРИ АНАЛИЗЕ БАЗЫ МНОГОМЕРНЫХ ДАННЫХ ПРИ РЕШЕНИИ ЗАДАЧИ ПРИНЯТИЯ РЕШЕНИЙ

В работе предложен метод устранения аномальных измерений (выбросов) для повышения уровня качества многомерных данных при статистических исследованиях. Такая проблема возникает, например, в теории принятия управленческих решений, поскольку при вычислении оценок параметров вероятностных распределений наличие в выборке аномальных (то есть таких, которые значительно увеличивают доверительный интервал) измерений способно исказить результаты статистического исследования, а, следовательно, и основной задачи. Особенность предложенного метода состоит в том, что он является комбинацией статистических и геометрического методов, а именно: метода оценки Гествирта, процедуры Тьюки и модифицированного метода построения выпуклой оболочки конечного множества точек многомерного пространства. Множеству многомерных данных ставится в соответствие множество точек многомерного пространства. Для нахождения и исключения выбросов строится последовательность вложенных выпуклых оболочек – n – политопов, каждый из которых описывается пересечением полупространств (опорных гиперграней). Приводится подробный алгоритм нахождения аномальных измерений. Их исключение соответствует последовательному исключению граничных точек вложенных выпуклых оболочек. Оценка Гествирта дает условие останова работы алгоритма. Предложенный метод не требует больших вычислительных затрат и может широко использоваться при решении как теоретических, так и практических задач, связанных с обработкой многомерных данных. Приведены численные результаты работы метода с количеством компонент данных 4 и 5.

Ключевые слова: статистическое исследование, теория принятия решений, аномальные измерения, выбросы, выпуклая оболочка, n – мерный политоп, многомерное пространство.

M. S. SOFRONOVA

METHOD FOR ELIMINATING ANOMALOUS MEASUREMENTS IN ANALYSIS OF THE MULTI-DIMENSIONAL DATABASE IN SOLVING THE DECISION-MAKING PROBLEM

The paper proposes a method for eliminating abnormal measurements (outliers) to improve the quality of multivariate data in statistical studies. Such a problem arises, for example, in the theory of managerial decision-making, since when calculating estimates of the parameters of probability distributions, the presence of anomalous (that is, those that significantly increase the confidence interval) measurements in the sample can distort the results of a statistical study, and, consequently, the main problem. The peculiarity of the proposed method is a combination of statistical and geometric methods, namely: the Gestwirt estimation method, the Tukey procedure, and a modification of the method for constructing the convex hull of a finite set of points in a multidimensional space. A set of multidimensional data is associated with a set of points of a multidimensional space. To find and eliminate outliers, a sequence of nested convex hulls, n – polytopes, is constructed, each of which is described by the intersection of half-spaces (support facets). A detailed algorithm for finding anomalous measurements is given. Their elimination corresponds to the successive elimination of the boundary points of nested convex hulls. The Gestwirt estimate gives the condition for stopping the operation of the algorithm. The proposed method does not require large computational costs and can be widely used in solving both theoretical and practical problems related to the processing of multidimensional data. The numerical results of the method with the number of data components 4 and 5 are presented.

Key words: statistical research, decision theory, anomalous measurements, outliers, convex hull, n – dimensional polytope, multidimensional space.

Вступ. Навколишнє середовище сучасного бізнесу пред'являє управлінцям все більше вимог і змушує приймати рішення все швидше. Тому важливою частиною процесу прийняття обґрунтованих управлінських рішень поряд з інтуїцією і досвідом менеджера сьогодні стають *статистичні методи обробки даних*, як традиційні (*аналіз динаміки*, *варіації, кореляційно-регресійні методи*), так і сучасні, що вимагають серйозної підготовки і застосування спеціальних програмних продуктів (*інтелектуальний аналіз даних*, *поглиблення даних* і *експериментальний дизайн*). Одним з моментів, що призводить до спотворення результатів статистичного дослідження і, відповідно, до помилкового рішення, є присутність в сукупності спостережень аномальних вимірю-

вань – *викидів*.

Виявлення викидів на різних етапах обробки даних дозволяє виявити помилки спостереження і, навіть, фіктивні дані. Поява аномальних вимірювань може бути наслідком помилок в даних (неточності вимірювання, округлення, невірного запису тощо), наявності шумових об'єктів (невірно класифікованих об'єктів), присутності об'єктів «інших» вибірок (наприклад, показаннями датчика, що зламався).

Зауважимо, що окрім теорії прийняття рішень, знаходження аномальних вимірювань є актуальним при виявленні підозрілих банківських операцій, нестандартних гравців на біржі, неполадок у механізмах за показаннями датчиків, в медичній діагностиці, сейсмології тощо. Тому для підвищення рівня якості аналітичних даних необхідно знизити вплив аномальних вимірювань (мінімізувати похибку результату) або зовсім виключити їх.

Аналіз останніх досліджень. Як показують статистичні дослідження [1], найрізноманітніші наукові, промислові, економічні та інші дані містять, як правило, 5 – 10 %, а іноді і більше, аномальних вимірювань, що істотно знижують ефективність застосування багатьох класичних процедур.

На теперішній час розроблено багато методів, що дозволяють мінімізувати похибку результату: це *робастні методи обробки*, *бутстреп методи* тощо [2 – 4]. У роботі [5] проведено аналіз існуючих методів робастного аналізу та бутстреп методів, наведено умови вибору найбільш ефективного для процедури аналізу даних, в залежності від прикладної задачі, що розв'язується.

Серед робастних важливий клас складають методи оцінювання, відомі як *методи Гествірта* [6]. Ці методи ґрунтуються на ідеї, згідно з якою найбільша довіра виявляється даним, ближчим до «центру» вибірки. Для реалізації цих методів можна використовувати геометричні поняття і методи, оскільки між геометрією і статистикою є тісний зв'язок, обумовлений тим, що багатовимірні статистичні дані можна розглядати як точки в евклідовому просторі.

У [6] запропоновано процедуру виключення припустимих аномальних вимірювань в одновимірному просторі, яка здійснюється наступним чином.

Нехай є N точок на прямій. Простий метод усунення припустимих викидів (тобто аномальних вимірювань) полягає у видаленні частини точок з лівого і правого боків – по $\lceil \alpha N \rceil$ точок з кожного боку (рис. 1). Середнє значення обчислюється за ($N - 2\lceil \alpha N \rceil$) точками, що залишилися (рис. 1).

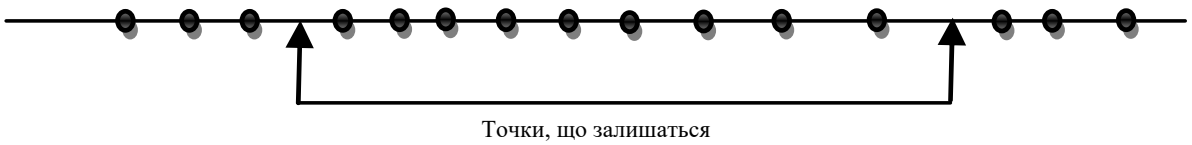

Рис. 1 – Виключення аномальних вимірювань в одновимірному просторі ($\alpha = 0{,}2$).

Зауважимо, що значення α задається, виходячи з певних теоретичних чи практичних міркувань, та задовольняє умові:

$$\begin{cases} N - 2\alpha N \geq n, \\ \alpha > 0. \end{cases} \Rightarrow \; 0 < \alpha \leq \frac{1}{2} - \frac{n}{2N}, \tag{1}$$

де n – вимірність простору.

Виникає питання: що представляє собою аналогічна процедура у випадку більш високої вимірності? *Тьюкі* [7] запропонував процедуру, відому під назвою *лущення*, що полягає у видаленні границі опуклої оболонки множини з подальшим видаленням границі опуклої оболонки множини, що залишилася, і так доти, доки не залишиться лише $\left(N - 2[\alpha N]\right)$ точок [8].

При побудові опуклої оболонки скінченної множини точок у R^n ($n \geq 3$) основна частина методів, що існують, наприклад, методи *загортання подарунка*, *під-над* [9] породжують повний опис межі (*граф граней*) опуклої оболонки. Як наслідок, процедура опису (пошуку) усіх *підграней* ускладнює й істотно впливає на часову складність методу. Проте побудова опуклої оболонки для задачі виключення аномальних вимірювань не потребує повного опису усіх підграней. Метод, запропонований у [10], дозволяє спростити процедуру побудови опуклої оболонки, завдяки пошуку лише її вершин та опорних *гіпергранеи*.

Враховуючи сказане вище, актуальною є **задача розробки методу усунення аномальних вимірювань** (даних), що базується на понятті опуклої оболонки скінченної множини точок багатовимірного простору, при аналізі багатовимірних даних (з кількістю компонент більше трьох).

Постановка задачі. Нехай на певному етапі розв'язання основної задачі одержано набір даних $A=(a_1, a_2, ..., a_m)$. Кожне a_j складається з n компонент a_{ji}: $a_j=(a_{j1}, a_{j2}, ..., a_{jn})$, $j \in J_m=\{1, 2, ..., m\}$, $i \in J_n=\{1,\ 2,\ ..., n\}$.

Необхідно проаналізувати наявні дані, виключивши ті, що призведуть до хибного результату, для одержання подальшого ефективного розв'язання основної (економічної, промислової тощо) задачі.

Викладення основного матеріалу дослідження. Для аналізу даних зручно використати статистичні методи. А саме, представити A як вибірку з варіантами a_j, $j \in J_m$, за якою і шукати оцінки параметрів ймовірнісних розподілів.

Для виявлення аномальних варіант запропонуємо комбінований метод, що базується на методі оцінювання Гествірта, процедурі Тьюкі та модифікації метода побудови опуклої оболонки скінченої множини точок багатовимірного простору, описаного в [10].

Поставимо у відповідність кожній варіанті a_j точку $A_j=(x_{j1}, x_{j2}, ..., x_{jn})$ $n-$вимірного евклідового простору, де $x_{ji}=a_{ji}$, $j \in J_m$, $i \in J_n$. Одержимо точкову множину $A=\{A_1, A_2, ..., A_m\} \subset R^n$ потужності m: $|A|=m$, де $m \geq n+1$.

Послідовність подальших дій:

1. Побудова опуклої оболонки $C_0=conv(A)$ точкової множини A з подальшим видаленням її граничних точок з множини A.

2. Формування множини A_l-відкоригованої множини A, $l=1, 2, ...$ Побудова $C_l=conv(A_l)$ та видалення її граничних точок з множини A_l.

3. Пункт 2 повторюється, доки у поточній множині A_l залишиться менше $(m-2[\alpha m])$ точок, а у попередній, A_{l-1} – не менше.

У результаті одержуємо послідовність $C_0 \supset C_1 \supset .. \supset C_{\tilde{s}}$ (рис. 2), у якій $C_{\tilde{s}}=conv(A_{\tilde{s}})$, а потужність результуючої множини $A_{\tilde{s}}$:

$$\left(|A_{\tilde{s}}| \geq m-2[\alpha m]\right) \wedge \left(|A_{\tilde{s}+1}| < m-2[\alpha m]\right).$$

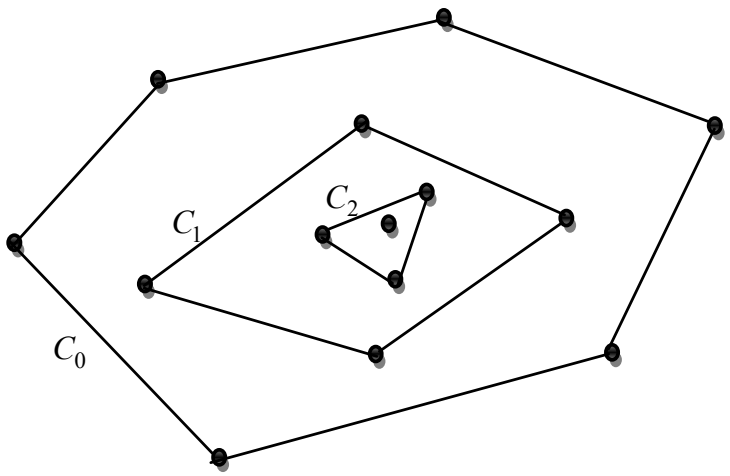

Рис. 2 – Приклад знаходження множини $A_{\tilde{s}}$ у двовимірному випадку ($\alpha=0,4$).
Елементи опуклої оболонки C_2 – це елементи шуканої множини $A_{\tilde{s}}$ ($\tilde{s}=2$).

Згідно з *теоремою McMullen*, *Shephard* [11], опукла оболонка скінченної множини точок у R^n є опуклим $n-$*вимірним політопом*, де під опуклим $n-$вимірним політопом ($n-$політопом) мається на увазі непорожня континуальна обмежена $n-$*вимірна поліедральна множина*, за умови, що ця множина не є підмножиною ніякого простору меншої вимірності [10]:

$$conv(A) \subset R^n,\ conv(A) \not\subset R^k,\ k \leq n-1.$$

Тому задача побудови опуклої оболонки множини точок можна звести до задачі побудови $n-$політопа на точках цієї множини, описаного набором орієнтованих гіперплощин.

Опишемо детальніше основні процедури побудови результуючої множини $A_{\tilde{s}}$. Припускаємо, що початкові дані задані коректно, тобто не всі точки A_j належать одній гіперплощині f (і можна побудувати початкову опуклу оболонку):

$$\overline{\exists} f : A_j \in f,\ j \in J_m. \tag{2}$$

Зауважимо, що, в основній задачі це означає незалежність компонент a_{ji} багатовимірних даних a_j, $j \in J_m$, $i \in J_n$, в наступному сенсі.

Назвемо компоненти $a_{j1}, a_{j2}, ..., a_{jn}$ незалежними [12], якщо виконується умова:

$$\overline{\exists} \varphi(x) : a_{ji_1} = \varphi(a_{ji_2}),\quad i_1 \neq i_2,\ i_1, i_2 \in J_n;\ j \in J_m,$$

де $\varphi(x)$ – деяка функція.

Якщо залежність спостерігається для деякої пари компонент з номерами, наприклад, q та z, по всіх даних a_j, $j = 1, 2, ..., m$, то можна виключити компоненту a_{jq} (або a_{jz}) з розгляду в усіх даних a_j. Це відповідає розгляду геометричної задачі у $(n-1)$ – вимірному просторі.

Процедура 1. Побудова опуклої оболонки C_0.

1.1. *Побудова початкового* $n-$*симплекса* S^1. Формуємо множину крайніх точок $K \subseteq A$. Зауважимо, що для точки $A'(x'_1, x'_2, ..., x'_n) \in K \subseteq A$ виконується умова:

$$\left(x'_k = \max_{j \in J_m} x_{jk}\right) \vee \left(x'_k = \min_{j \in J_m} x_{jk}\right),\ k \in \{1,\ 2, ..., n\}; \tag{3}$$

Якщо існує декілька точок, для яких виконується умова (3), тобто

$$\exists\ m_1, m_2 \in J_m : x'_k = x_{m_1 k} = x_{m_2 k},\ k \in \{1, 2, ..., n\} \Rightarrow \begin{bmatrix} x'_k = x_{m_1 k},\ k_1 \geq k_2; \\ x'_k = x_{m_2 k},\ k_1 < k_2, \end{bmatrix}$$

де $k_1(k_2)$ – кількість осей k, $k \in \{1, 2, ..., n\}$, на яких для $x_{m_1 k}$ $\left(x_{m_2 k}\right)$ виконується умова (3).

Нехай $|K| = \tilde{k}$.

При $\tilde{k} \geq n$ будуємо на n точках $A'_1, A'_2, ..., A'_n$ множини K гіперплощину $f^0 : \sum_{k=1}^{n} a_k x_k + a_0 = 0$ – одну з $C_{\tilde{k}}^n$ гіперплощин, де $C_{\tilde{k}}^n = \dfrac{\tilde{k}!}{n!(\tilde{k}-n)!}$ – число комбінацій з $\tilde{k}$ елементів по n.

При $\tilde{k} < n$ будуємо на n точках множини A гіперплощину $f^0 : \sum_{k=1}^{n} a_k x_k + a_0 = 0$ – одну з C_m^n гіперплощин.

Шукаємо точку $A_0\left(x_{01}, x_{02}, ..., x_{0n}\right) \in A$, для якої:

$$\left(\left|\delta_{A_0}(f^0)\right| = \left|f^0(A_0)\right| = \max_{\substack{A_i \in A \\ i \in J_m}} \left\{\left|f^0(A_i)\right|\right\}\right) \wedge \left(\left|\delta_{A_0}(f^0)\right| \neq 0\right).$$

Зауважимо, що в силу припущення (2) така точка існує.

Формуємо точкову множину $\tilde{A} = \{\tilde{A}_1, \tilde{A}_2, ..., \tilde{A}_{n+1}\} = \{A'_1, A'_2, ..., A'_2, A_0\}$ і будуємо $conv(\tilde{A})$ – $n-$*симплекс* S^1. Для цього:

– з точок множини $\tilde{A} = \{\tilde{A}_j\}_{j=1}^{n+1}$ генеруємо $(n+1)$ наборів по n точок $\tilde{A}_1^t, \tilde{A}_2^t, ..., \tilde{A}_n^t \in \tilde{A}$, $t \in J_t = \{1,\ 2, ..., n+1\}$;

– будуємо гіперплощини $f_t^1 :\ \sum_{k=1}^{n} a_k^t x_k + a_0^t = 0$, $t \in J_t$. Зауважимо, що

$$\exists\, t \in \{1, 2, ..., n+1\} : f_t^1 = f^0.$$

– орієнтуємо недодатньо відносно множини $\tilde{A}$ гіперплощини f_t^1, $t \in J_t$:

– обираємо точку $A''(x_1'', x_2'', ..., x_n'') \in \tilde{A} : f_t^1(A'') \neq 0$;

– обчислюємо величину $\delta_{A''}(f_t^1) = \sum_{k=1}^{n} a_k^t x_k'' + a_0^t$;

– якщо $\delta_{A''}(f_t^1) > 0$, то у рівнянні гіперплощини f_t^1 змінюються знаки всіх коефіцієнтів і вільного члена на протилежні, тобто $a_k^t = -a_k^t$, $k = 0, 1, ..., n$.

Таким чином, одержимо $conv(\tilde{A})$ – n – симплекс S^1, межа якої описується системою нерівностей:

$$\begin{cases} \sum_{k=1}^{n} a_k^1 x_k + a_0^1 \le 0; \\ \sum_{k=1}^{n} a_k^2 x_k + a_0^2 \le 0; \\ ... \\ \sum_{k=1}^{n} a_k^{n+1} x_k + a_0^{n+1} \le 0. \end{cases}$$

1.2. *Коригування множини A (виключення з A внутрішніх точок та тих, що є граничними, але не є вершинами n – політопа S^h, $h = 1, 2, ...$).*

Позначимо через $B^h \subset A$ – множину точок після коригування множини A на h – му кроці; A^h – множина вершин n – політопа S^h, $h = 1, 2, ...$, причому $A^1 = \tilde{A}$; $H_{S^h} = \{f_t^h\}_{t=1}^{t_h}$ – множина гіперплощин f_t^h, $t \in J_{t_h} = \{1, 2, ..., t_h\}$, що визначає границю n – політопа S^h (причому $t_1 = n+1$).

Формуємо B^h, для цього виключаємо з розгляду точку $A'''(x_1''', x_2''', ..., x_n''') \in A$, якщо

$$(A''' \in B^{h-1}) \wedge (A''' \in S^h \setminus A^h) \text{ (при } B^0 = A).$$

Тобто, якщо:

– $\delta_{A'''}(f_t^h) < 0$, $t \in J_{t_h}$ (A''' – внутрішня точка n – політопа) або

– $\left(\exists\, \tilde{f}_1^h, \tilde{f}_2^h, ..., \tilde{f}_{l_0}^h \in H_{S^h} : \ \delta_{A'''}(\tilde{f}_l^h) = 0, \ l \in \{1, 2, ..., l_0\}\right) \wedge$

$$\wedge \left(\forall \tilde{\tilde{f}}_p^h \in H_{S^h} \setminus \left\{\tilde{f}_l^h\right\}_{l=1}^{l_0} : \ \delta_{A'''}(\tilde{\tilde{f}}_p^h) < 0, \ p = 1, 2, ..., t_h - l_0 \right)$$

(A''' – гранична, але не є вершиною n – політопа).

Тоді множина $B^h = B^{h-1} \setminus \left\{A_r'''\right\}_{r \in \mathrm{N}}$, $h = 1, 2, ...$, де r – число точок виду A'''.

1.3. *Побудова n – політопа (опуклої оболонки) S^{h+1}, $h = 1, 2, ...$.*

Для кожної гіперплощини $f_t^h \in H_{S^h}$, $t \in J_{t_h}$, обираємо точку $A_0(x_{01}, x_{02}, ..., x_{0n}) \in B^h$ з максимальним відхиленням від площини f_t^h, тобто

$$\delta_{A_0}(f_t^h) = a_1^t x_{01} + a_2^t x_{02} + ... + a_n^t x_{0n} + a_0^t = \max_{\substack{A_i \in B^h, \\ i \in \{1, 2, ..., m\}}} \left\{a_1^t x_{i1} + a_2^t x_{i2} + ... + a_n^t x_{in} + a_0^t\right\}.$$

Можливі такі випадки:

– $\delta_{A_0}(f_t^h) < 0 \ \Rightarrow \ f_t^h \in H_{conv(A)}$, де $H_{conv(A)}$ – множина опорних гіперплощин (всі точки множини A лежать по один бік від гіперплощини). Позначимо через $P_{conv(A)}$ – множину точок, на яких побудовано опорні гіперплощини;

– $\left(\delta_{A_0}(f_t^h) = 0\right) \wedge$

$$\wedge \left(\left(\exists f_t^* \in H_{S^h}, \ t = 1, 2, ..., t^* (t^* < n) : \ \delta_{A_0}(f_t^*) = 0 \right) \wedge \left(\forall f_t^{**} \in H_{S^h}, \ t = 1, 2, ..., \ t_h - t^* : \ \delta_{A_0}(f_t^{**}) < 0 \right) \right)$$

(A_0 належить гіперграні, через яку проходить f_t^h). В цьому випадку необхідно перейти до розгляду наступної гіперплощини та повторити спочатку крок 1.3 з обрання точки A_0;

$$- \left(\left(\delta_{A_0}(f_t^h) > 0\right) \vee \left(\delta_{A_0}(f_t^h) = 0\right)\right) \wedge \left(\exists\ f_t^* \in H_{S^h},\ t \in \{1, 2, ..., t_h\} :\ \delta_{A_0}(f_t^*) > 0\right) \Rightarrow$$

$$\Rightarrow \exists\ \left\{f_s^*\right\}_1^{g-1} \in H_{S_h},\ g-1 < t_h : \delta_{A_0}(f_s^*) > 0,\ s = 1, 2, ..., g-1,\ f_s^* \neq f_t^h\ (A_0 \text{ лежить поза } S^h).$$

Введемо у розгляд множини $H_{A_0} = \{f_1^*, ..., f_{g-1}^*, f_g^*\}$, де $f_g^* = f_t^h$, та $P_{A_0} = \bigcup_{s=1}^{g} P_{A_0}^s$, де $P_{A_0}^s$ – множина точок з A^h (вершин $n-$політопа S^h), через які проходить f_s^*, $s = 1, 2, ..., g$. Нехай $\left|P_{A_0}^s\right| = g_s'$, $\left|P_{A_0}\right| = g'$, тоді $\sum_{s=1}^{g} g_s' \geq g'$.

Побудуємо допоміжні $n-$симплекси $\hat{S}_d$, $d \in J_\mu = 1, 2, ..., \mu$, де $\mu = \sum_{s=1}^{g} C_{g_s'}^n$. Кожен $n-$симплекс $\hat{S}_d$ будуємо, використовуючи процедуру 1.1, на $(n+1)-$й точці множини $\hat{A}_d = \{\hat{A}_{d1}, \hat{A}_{d2}, ..., \hat{A}_{d(n+1)}\}$, елементи якої обираються за правилом: $\hat{A}_{d1}, \hat{A}_{d2}, ..., \hat{A}_{dn} \in P_{A_0}^s$, $s \in \{1, 2, ..., g\}$, $\hat{A}_{d(n+1)} = A_0$, $d \in J_\mu$. Нехай $\hat{S} = \bigcup_{d=1}^{\mu} \hat{S}_d$, $\{\hat{f}_t^d\}_{t=1}^{n+1}$ – опорні гіперплощини $n-$симплекса $\hat{S}_d$, $d \in J_\mu$; $H_0 \subset \left\{\hat{f}_1^1, ..., \hat{f}_{n+1}^1, \hat{f}_1^2, ..., \hat{f}_{n+1}^2, ..., \hat{f}_1^\mu, \hat{f}_{n+1}^\mu\right\}$, де H_0 – множина гіперплощин, з якої вилучені ті, які не є опорними для множини точок $\{A_0\} \cup A^h$. Зауважимо, що $\left|H_0\right| < (n+1)\mu$.

Формуємо множину гіперплощин $H_{S^{h+1}} = H_{S^h} \cup H_0$, $\left|H_{S^{h+1}}\right| = t_{h+1}$, які будуть опорними для $n-$політопа S^{h+1}. Зауважимо, що потрібно врахувати випадок співпадіння декількох площин.

Використовуючи процедуру 1.2, сформуємо множину точок A^{h+1}, елементами якої є вершини $n-$політопа S^{h+1}.

За скінченне число ітерацій за умови, що $A^{\tilde{h}} = B^{\tilde{h}}, \tilde{h} \in \mathrm{N}$, (не існує жодної зовнішньої по відношенню до $S^{\tilde{h}}$ точки множини A) одержимо шукану опуклу оболонку ($n-$політоп) C_0, що задається набором опорних гіперплощин (елементи $H_{conv(A)}$) та вершин (елементи $P_{conv(A)}$).

Зауважимо, що верхньою оцінкою $f(n,m)$ числа опорних гіперплощин буде [9]:

$$f(n,m) = \begin{cases} \dfrac{2m}{n} C_{m-\frac{n}{2}-1}^{\frac{n}{2}-1} & \text{для парних } n; \\ 2C_{m-\left\lfloor\frac{n}{2}\right\rfloor-1}^{\left\lfloor\frac{n}{2}\right\rfloor} & \text{для непарних } n. \end{cases}$$

Процедура 2. Видалення граничних точок опуклої оболонки C_0 (множини $\gamma_0 = fr\, C_0$) з множини A. Формування множини A_1 – відкоригованої множини A.

У результаті використання процедури 1 одержимо:

– множину $D_0 = \left\{V_1^0, V_2^0, ..., V_{d_0}^0\right\} \subset \gamma_0$ $\left(D_0 = P_{conv(A)}\right)$ граничних точок $V_t^0(x_{t1}^0, x_{t2}^0, ..., x_{tn}^0)$, $t = 1, 2, ..., d_0$, опуклої оболонки C_0, $D_0 \subseteq A$, $\left|D_0\right| = d_0$, $d_0 \leq m$;

– набір з h_0 орієнтованих (недодатньо) опорних гіперплощин $H_{C_0} = \{f_1^0, f_2^0, ..., f_{h_0}^0\}$, $\left(H_{C_0} = H_{conv(A)}\right)$.

Для значення d_0 потужності множини D_0 перевіряємо виконання умови:

$$m - d_0 < m - 2\left[\alpha m\right],$$

тобто

$$d_0 > 2\left[\alpha m\right]. \tag{4}$$

Якщо умова (4) виконується, то це означає, що A і є шуканою множиною $A_{\tilde{s}}$.

В іншому випадку, видаляємо з множини A ті точки, що є елементами множини D_0 та формуємо точкову множину $A_1 = A \setminus D_0$ потужності $d_1 = m - d_0$.

Покладемо $l = 1$.

Процедура 3. Побудова опуклої оболонки C_l та видалення її граничних точок з множини A_l.

Для побудови $C_l = conv(A_l)$ використаємо процедуру 1 (в якості множини A будемо розглядати множину A_l). В результаті одержимо:

– множину $D_l = \left\{V_1^l, V_2^l, ..., V_{d_l}^l\right\} \subset \gamma_l$ (де $\gamma_l = frC_l$) граничних точок $V_t^l(x_{t1}^l, x_{t2}^l, ..., x_{tn}^l)$, $t = 1, 2, ..., d_l$, опуклої оболонки C_l, $D_l \subseteq A_l$, $|D_l| = d_l$, $d_l < d_{l-1}$;

– набір з h_l орієнтованих (недодатньо) опорних гіперплощин $H_{C_l} = \{f_1^l, f_2^l, ..., f_{h_l}^l\}$.

Для значення потужності d_l множини D_l перевіряємо виконання умови:

$$m - \sum_{k=0}^{l} d_k < m - 2[\alpha m],$$

тобто

$$\sum_{k=0}^{l} d_k > 2[\alpha m]. \qquad (5)$$

Якщо умова (5) виконується, то це означає, що $A_{\tilde{s}} = A_{l-1}$.

В іншому випадку, видаляємо з множини A_l ті точки, що є елементами множини D_l.

Покладемо $l = l + 1$.

Формуємо точкову множину $A_l = A_{l-1} \setminus D_{l-1}$ потужності $|A_l| = m - \sum_{k=0}^{l-1} d_k$.

Повторюємо процедуру 3 до виконання для поточного l умови (5).

В результаті одержимо $C_{\tilde{s}}$ – опуклу оболонку множини $A_{\tilde{s}}$, що містить $d_{\tilde{s}}$ точок $V_t^{\tilde{s}}$, $t = 1, 2, ..., d_{\tilde{s}}$, множини A.

Кожній точці $V_t^{\tilde{s}}(x_{t1}^{\tilde{s}}, x_{t2}^{\tilde{s}}, ..., x_{tn}^{\tilde{s}})$, $t = 1, 2, ..., d_{\tilde{s}}$, відповідає варіанта $a'_t = (a'_{t1}, a'_{t2}, ..., a'_{tn})$ початкового набору даних A. За одержаним набором варіант $a'_1, a'_2, ..., a'_{d_{\tilde{s}}}$ тепер можна шукати оцінки параметрів ймовірнісного розподілу. Причому, варіанти, що могли призвести до хибного результату, завдяки запропонованому методу були виключені з розгляду.

Результати роботи методу. Результати проведення чисельних експериментів надано в табл. 1, де m – кількість даних (початкова кількість варіант); n – кількість компонент; $d_{\tilde{s}}$ – кількість варіант, що залишилися після виключення аномальних; $\alpha \in \{0{,}1; 0{,}15; 0{,}2\}$. Зауважимо, що значення компонент $(a'_{i1}, a'_{i2}, ..., a'_{in})$, $t = 1, 2, ..., d_{\tilde{s}}$, обираються випадковим чином за умови, що $a'_{ti} \in N$, $a'_{ti} \in [1, 100]$, $i \in J_n$.

Таблиця 1 – Результати використання розробленого методу

n	m	$d_{\tilde{s}}$	α	n	m	$d_{\tilde{s}}$	α
4	30	24	0,1	5	30	22	0,1
	40	32	0,2		40	27	0,2
	50	37	0,15		50	41	0,15

Як бачимо з табл. 1, після корегування початкових даних, тобто усунення аномальних вимірювань за допомогою розробленого метода, кількість варіант зменшилась в середньому на 24 % в залежності від кількості компонент даних. Подальше обчислення базових характеристик (зокрема, середнє, розмах та коефіцієнт варіації, дисперсія тощо) здійснюється на відкоригованій вибірці.

Перспективи подальших досліджень. Перспективним є дослідження корекції впливу викидів. Методи корекції впливу викидів можуть знижувати вплив викидів на результат аналізу даних без видалення значень, які розпізнані як викиди. Альтернативний спосіб полягає в проведенні аналізу двічі: при наявності викидів і без викидів.

Висновки. У статті запропоновано комбінований метод усунення аномальних вимірювань при аналізі багатовимірних даних під час розв'язання задачі прийняття рішень, що базується на використанні методу оцінювання Гествірта, процедури Тьюкі та модифікації метода побудови опуклої оболонки скінченої множини точок ба-

гатовимірного простору. Головні особливості розробленого методу полягають у наступному: початкова задача зводиться до статистичної, яка в свою чергу – до геометричної (дані – варіанти – точки); знаходження та виключення аномальних варіант відбувається через побудову послідовності вкладених опуклих оболонок – n – політопів; кожен n – політоп описується набором граничних точок та гіперплощин, що зменшує часову складність знаходження n – політопа (а значить, і опуклої оболонки); обчислення оцінок параметрів ймовірнісного розподілу (тобто проведення аналізу даних) здійснюється на варіантах, що відсортували.

Таким чином, для підготовки обґрунтованого управлінського рішення можна при коригуванні даних на основі статистичного процесу використати описаний алгоритм виявлення викидів, а потім вже обирати найбільш раціональне рішення. Звичайно, можна обрати варіант рішення і без аналізу викидів. Однак наявність у сукупності аномальних спостережень може привести до використання неправдивих передумов при прийнятті рішення. Ціна помилки в результаті некомпетентного або недостатньо підготовленого рішення буває часом достатньо висока.

Список літератури

1. *Hampel F. R.* Robust estimation: a condensed partial survey // Z. Wahrscheinlichkeits – Theorie and Verw. Geb. – 1973. – 27. – pp. 87 – 104.
2. *Поляк Б. Т.*, *Щербаков П. С.* Робастная устойчивость и управление. – М. : Наука, 2002. – 303 с.
3. *Шуленин В. П.* Математическая статистика. Ч. 3: Робастная статистика. – Томск : Изд-во НТЛ, 2012. – 520 с.
4. *Эфрон Б.* Нетрадиционные методы многомерного статистического анализа. – М. : Финансы и статистика, 1988. – 261 с..
5. *Гожий А. П.*, *Коваленко И. И.* Системное использование робастных и бутстреп методов в задачах анализа данных. – АСАУ – 9(29). – 2006. – С. 38 – 49.
6. *Gastwirth J.* On robust procedures // J. Amer. Stat. Assn. – 1966. – pp. 929 – 948.
7. *Tukey J. W.* A survey of sampling from contaminated distributions // Contributions to probability and statistics. – 1960. – Vol. 2. – pp. 448 – 485.
8. *Huber P. J.* Robust statistics : A revier. // Ann. Math. Stat. – 1972. – 43(3). – pp. 1041 – 1067.
9. *Препарата Ф.*, *Шеймос М.* Вычислительная геометрия. Введение. – М. : Мир, 1989. – 478 с.
10. *Гиль Н. И.*, *Софронова М. С.* Об одном подходе к построению выпуклой оболочки конечного множества точек в R^n // Штучний інтелект. – 2009. – № 4 – С. 30 – 36.
11. *McMullen P.*, *Shephard G*. Convex Polytopes and the Upper Bound Conjecture. – Cambridge: University Press, 1971.
12. *Погожих М. І.*, *Софронова М. С.* Математичне моделювання задач оптимізації в економіці // Економічна стратегія і перспективи розвитку сфери торгівлі та послуг: зб. наук. пр. ХДУХТ. – Х. : ХДУХТ, 2017. – Вип.1 (25). – С. 121 – 131.

References (transliterated)

1. Hampel F. R. Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeits – Theorie and Verw. Geb*. 1973, no. 27, pp. 87–104.
2. Polyak B. T., Shcherbakov P. S. *Robastnaya ustoychivost' i upravlenie* [Robust stability and control]. Moscow, 2002. 303 p.
3. Shulenin V. P. *Matematicheskaya statistika. CH. 3: Robastnaya statistika* [Mathematical statistics. Part 3: Robust statistics]. Tomsk, 2012. 520 p.
4. Efron B. *Netraditsyonnye metody mnogomernogo statisticheskogo analiza* [Non-traditional methods of multivariate statistical analysis]. Moscow, Finance and statistics Publ., 1988. 261 p.
5. Gozhiy A. P., Kovalenko I. I. *Sistemnoe ispol'zovanie robastnykh i butstrep metodov v zadachakh analiza dannykh* [Systematic use of robust and bootstrap methods in data analysis problems]. ASAU, 2006, no. 9 (29), pp. 38–49.
6. Gastwirth J. On robust procedures. *J. Amer. Stat. Assn*. 1966, pp. 929–948.
7. Tukey J. W. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*. 1960, vol. 2, pp. 448–485.
8. Huber P. J. Robust statistics : A revier. *Ann. Math. Stat*. 1972, no. 43(3), pp. 1041–1067.
9. Preparata F., Sheimos M. *Vychislitel'naya geometriya. Vvedenie* [Computational geometry. Introduction]. Moscow, 1989. 478 p.
10. Gil N. I., Sofronova M. S. Ob odnom podkhode k postroeniyu vypukloy obolochki konechnogo mnozhestva tochek v R^n [On an approach to the construction of the convex hull of a finite set of points in R^n]. *Shtuchniy intelekt* [Artificial intelligence]. 2009, no. 4, pp. 30–36.
11. McMullen P., Shephard G. *Convex Polytopes and the Upper Bound Conjecture.* Cambridge, University Press, 1971.
12. Pogozhikh M. I., Sofronova M. S. Matematychne modelyuvannya zadach optimizatsiyi v ekonomitsi [Mathematical modeling of optimization problems in economics]. *Ekonomichna strategiya i perspektyvy rozvytku sfery torgivli ta poslug* [Economic strategy and prospects for the development of the sphere of trade and services]. Kharkiv, 2017, vol. 1 (25), pp. 121–131.

Надійшла (received) 01.10.2021

Відомості про авторів / Сведения об авторах / Information about authors

Софронова Марина Сергіївна – кандидат фізико-математичних наук, доцент, доцент кафедри вищої математики, Національний технічний університет «Харківський політехнічний інститут», м. Харків; тел.: (057) 707-60-87; e-mail: m_myravyova@ukr.net.

Софронова Марина Сергеевна – кандидат физико-математических наук, доцент, доцент кафедры высшей математики, Национальный технический университет «Харьковский политехнический институт», г. Харьков; тел.: (057) 707-60-87; e-mail: m_myravyova@ukr.net.

Sofronova Maryna Sergeevna – Candidate of Physics and Mathematics Sciences, Associate Professor, Associate Professor at the Department of Higher Mathematics, National Technical University «Kharkiv Polytechnic Institute», Kharkiv; tel.: (057)707-60-87; e-mail: m_myravyova@ukr.net.