УДК 31.311.2; 629.3.073
[1]Ханін О.Г. к.ф.-м.н., [2]Лотиш В.В., к.т.н, [2]Гуменюк П.О., к.т.н, [2]Гуменюк Л.О. к.т.н.
[1]Східноєвропейський національний університет імені Лесі Українки.
[2]Луцький національний технічний університет.

УДОСКОНАЛЕНИЙ МЕТОД χ^2-КЛАСТЕРИЗАЦІЇ ТА ЙОГО ЗАСТОСУВАННЯ ДО АНАЛІЗУ АВАРІЙНОСТІ НА АВТОМОБІЛЬНОМУ ТРАНСПОРТІ

Ханін О.Г., Лотиш В.В., Гуменюк П.О., Гуменюк Л.О. Удосконалений метод χ^2-кластеризації та його застосування до аналізу аварійності на автомобільному транспорті. Існує чимало методів кластеризації даних, але вони мають ряд недоліків, зокрема, з одного боку, це – неоднозначність розбиття масиву даних на групи, а з іншого – неможливість оцінити ступінь однорідності об'єктів, що належать одному і тому ж кластеру. Мета цієї роботи - розробити метод кластерного аналізу багатовимірних даних різної природи, який забезпечить однозначність розбиття набору незалежних вибірок на кластери, що не перетинаються, так, щоб ймовірність помилкової кластеризації не перевищувала певного наперед заданого рівня. Метод кластеризації ґрунтується на використанні критерію узгодженості χ^2 .

З іншої сторони, проблема аварійності на автомобільному транспорті є достатньо гострою у вітчизняних реаліях, оскільки рівень ДТП з потерпілими значно перевищує середній європейський. В той же час, кластеризація регіонів України за видами дорожньо-транспортних пригод з постраждалими, їх причинами та винуватцями дозволить зрозуміти спільні регіональні фактори, що впливають на рівень аварійності, визначити та впровадити кращі практики її запобігання. Саме тому представляється актуальним застосування запропонованого методу до аналізу аварійності на автомобільному транспорті в Україні.Розроблений метод реалізований програмно та застосований до порівняльного аналізу рівня аварійності на автомобільному транспорті по регіонах України.

Ключові слова: кластерний аналіз, біноміальний розподіл, довірчий інтервал, критерій узгодженості χ^2, перевірка статистичних гіпотез, похибка кластеризації, дорожньо-транспортні пригоди з постраждалими.

Ханин А.Г., Лотыш В.В., Гуменюк П.А., Гуменюк Л.А. Усовершенствованный метод χ^2 - кластеризации и его применение к анализу аварийности на автомобильном транспорте. Существует немало методов кластеризации данных, но они страдают, с одной стороны, неоднозначностью разбиения массива данных на группы, а с другой, не дают возможности оценить степень однородности объектов, принадлежащих одному и тому же кластеру. Цель этой работы - разработать метод кластерного анализа многомерных данных различной природы, который обеспечит однозначность разбиения набора независимых выборок на непересекающиеся кластеры, так, чтобы вероятность ложной кластеризации не превышала определенного заранее заданного уровня. Метод кластеризации основывается на использовании критерия согласия χ^2.

С другой стороны, проблема аварийности на автомобильном транспорте является достаточно острой в отечественных реалиях, поскольку уровень ДТП с пострадавшими значительно превышает средний европейский. В то же время, кластеризация регионов Украины по видам дорожно-транспортных происшествий с пострадавшими, их причинами и виновниками позволит понять общие региональные факторы, влияющие на уровень аварийности, определить и внедрить лучшие практики ее предотвращения. Именно поэтому представляется актуальным применение предложенного метода к анализу аварийности на автомобильном транспорте в Украине.Разработанный метод реализован программно и применен к сравнительному анализу уровня аварийности на автомобильном транспорте по регионам Украины.

Ключевые слова: кластерный анализ, биномиальное распределение, доверительный интервал, критерий согласия χ^2, проверка статистических гипотез, погрешность кластеризации, дорожно-транспортные происшествия с пострадавшими.

O. Khanin, V. Lotysh, P. Gumeniuk, L. Gumeniuk. Improved method of χ^2 - clusterization and its application to the analysis of emergency in automobile transport. There is a large variety of methods for data clustering, but all of them have numerous defects. On the one hand, it is the ambiguity of splitting the data array into groups, and on the other the impossibility to assess the degree of homogeneity of objects belonging to the same cluster. The purpose of this work is to develop the method for cluster analysis of multidimensional data of a different nature, which will ensure unambiguity of splitting a set of independent samples into clusters that do not intersect, so that the probability of false clustering does not exceed a certain predetermined level. The clustering method is based on the use of the χ^2 consistency criterion.

Besides, the problem of accidents on road transport is quite acute in the domestic realities, since the level of accidents with victims significantly exceeds the average European. At that, the clustering of the regions of Ukraine according to the types of road accidents with the victims, their causes and culprits will help to understand the general regional factors affecting the accident rate, to identify and implement the best practices for its prevention. That is why the application of the proposed method to the analysis of accidents on road transport in Ukraine is quite relevant. The developed method is implemented programmatically and applied to a comparative analysis of the accident rate in road transport in the regions of Ukraine.

Keywords: cluster analysis, binomial distribution, confidence interval, χ^2 consistency criterion, verification of statistical hypotheses, clustering error, road traffic accidents with victims.

Постановка проблеми. Кластеризація, тобто розбиття даних на однорідні, в певному розумінні, групи, представляє собою поширений метод аналізу даних шляхом зменшення розмірності їх масиву, допомагає виявити спільні та відмінні риси об'єктів дослідження [1]. Цей метод знайшов застосування при розгляді багатовимірних систем в техніці, економіці, фінансах, маркетингу, соціології, психології, медицині, тощо. Однак, головними його недоліками є, як правило, неоднозначність розбиття залежно від вибору першого (базового) об'єкту, з якого починається процес кластеризації, та відсутність міри якості кластеризації даних. Тому, головним чином, цей метод використовується як метод попереднього, розвідувального аналізу, який дозволяє на якісному рівні побачити певні закономірності і сформулювати гіпотези, що потребують подальшого

дослідження. Задачею авторів була побудова такого методу кластерного аналізу, який би, по можливості, був позбавлений цих недоліків, тобто відрізнявся однозначністю кластеризації та забезпечував певну наперед задану ймовірність її похибки. З іншого боку, оскільки проблема аварійності на автомобільному транспорті в Україні в останні роки є надзвичайно гострою [2 - 4] та її рівень суттєво перевищує середньоєвропейський [4], представляє інтерес застосування запропонованого алгоритму до регіонального аналізу причин, видів та винуватців ДТП з постраждалими на автотранспорті в Україні.

Метою роботи було, грунтуючись на методології порівняння емпіричних розподілів за критерієм узгодженості χ^2 [5], запропонувати та обґрунтувати удосконалений метод кластерного аналізу, який би забезпечив однозначність розбиття набору незалежних вибірок на кластери, що не перетинаються, так, щоб ймовірності помилкового віднесення вибірок до різних кластерів не перевищувала певного наперед заданого рівня; за допомогою розглянутого методу провести кластерний аналіз регіонів України за структурою ДТП з постраждалими на автомобільному транспорті за причинами, видами та винуватцями.

Запропонований метод кластерного аналізу ґрунтується на критерії узгодженості χ^2 [5], а ідея його застосування до задач кластерного аналізу в галузі маркетингу була представлена у роботі [6]. Однак, як і решта методів кластерного аналізу, цей метод мав неоднозначність результуючого розбиття, а також відсутність конкретної оцінки ймовірності помилкового віднесення до різних кластерів однорідних об'єктів. В даній роботі представлено удосконалений метод кластеризації багатовимірних даних, який дає однозначні результати, а також відповідні оцінки похибки кластеризації. На основі цього алгоритму в середовищі Delphi створена програма, яка застосована до аналізу значного масиву даних, пов'язаних з рівнем аварійності на автомобільному транспорті по регіонах України.

Аналіз досліджень. Як відомо, кластеризація – це процес розбиття заданої вибірки об'єктів на підмножини, що не перетинаються, які називаються кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів суттєво відрізнялися [7]. Термін «кластерний аналіз» уперше ввів Трайон (Tryon) [8] у 1939 році. Відтоді розроблено чимало методів кластерного аналізу, певну класифікацію яких наведено у багатьох роботах, зокрема [2,9-11]. Згідно цієї класифікації, запропонований нами метод відноситься до ієрархічних методів дівізівної кластеризації (Divisive Methods) із чітким багатоетапним алгоритмом. Заздалегідь невідомо, на скільки кластерів буде розбита сукупність багатовимірних даних. Вхідні дані для застосування запропонованого методу кластеризації повинні бути представлені у вигляді таблиці «Ознака - Кількість її спостережень».

Такий авторитет у галузі аналізу даних, як Тьюкі (Tukey) [12] поділяє статистичний аналіз на два етапи: розвідувальний та підтверджуючий. Перший етап включає перетворення даних спостережень і способи їх наочного представлення, що дозволяє виявити внутрішні закономірності, які проявляються в даних, тобто, фактично, сформулювати певні гіпотези. На другому етапі застосовуються традиційні статистичні методи оцінки параметрів і перевірки гіпотез. Загальноприйняті методи кластеризації є потужними засобами саме розвідувального аналізу [13]. Запропонований нами метод відрізняється тим, що забезпечує певну однозначність процесу кластеризації, і поєднує у собі обидва етапи. Він, завдяки розбиттю даних на підмножини, які складаються з багатовимірних даних, із ймовірнісними розподілами, схожими між собою за критерієм узгодженості Пірсона, надає інформацію для подальшого формулювання гіпотез щодо змістовних причин схожості чи відмінності даних, та одночасно дає можливість оцінити ймовірність помилкового віднесення «схожих» об'єктів до різних кластерів, тобто зробити певні статистичні висновки.

Однією з цілей кластеризації є стиснення великих обсягів даних: замість дослідження всього їх неосяжного масиву можна розглядати та порівнювати між собою по одному типовому представнику від кожного кластеру [7]. Безумовно, це дає можливість виявити глибинні причини, що розділяють представників різних кластерів. Оскільки нас цікавили саме такі причини регіональних відмінностей рівня аварійності на автомобільному транспорті в Україні, ми вирішили застосувати запропонований нами метод кластерного аналізу до цієї актуальної задачі.

Методологія досліджень. Припустимо, що спостерігаються m незалежних вибірок об'ємів n1, n2, …, nm, відповідно, з генеральних сукупностей, елементи яких приймають одне з r можливих значень. В якості таких значень можуть виступати певні якісні ознаки (групи факторів), наприклад, групи причин скоєння ДТП, або належність до певного інтервалу для кількісних ознак (в останньому випадку значення кількісних ознак повинні бути розбиті на r інтервалів, що не перетинаються). Нехай для кожної спостереженої вибірки побудовано емпіричний розподіл частот по групах. Наприклад, в таблиці 1 наведені розподіли частот по п'ятьох групах факторів, отримані за чотирма вибірками.

В один кластер хотілося б об'єднати ті вибірки, в яких співпадають теоретичні розподіли, тобто розподіли генеральних сукупностей, яким вони належать. Однак, на практиці теоретичні розподіли невідомі. Методологія попарного порівняння емпіричних розподілів за допомогою критерію χ^2 була розглянута нами в [5], але вона призводить до неоднозначного результату кластерного аналізу залежно від вибору еталонного розподілу на кожному кроці кластеризації.

Таблиця 1. Розподіл вибіркових частот по групах

Групи факторів або інтервали значень		Група 1	Група 2	Група 3	Група 4	Група 5	РАЗОМ
Вибірка 1	Кількість спостережень, що належать даній групі	4	103	24	90	32	283
Вибірка 2	Кількість спостережень, що належать даній групі	190	174	73	126	118	681
Вибірка 3	Кількість спостережень, що належать даній групі	211	82	79	77	122	561
Вибірка 4	Кількість спостережень, що належать даній групі	225	73	42	25	110	475

Модифікуємо запропонований у роботі [6] алгоритм кластеризації так, щоб на кожному новому кроці кластеризації в якості еталонної вибірки обиралася вибірка, найбільша за об'ємом. Розподіл цієї вибірки будемо називати еталонним вибірковим розподілом. Відповідно, генеральну сукупність, з якої узята ця вибірка будемо також називати еталонною, а її розподіл - еталонним теоретичним розподілом. Так, в нашому прикладі на першому кроці в якості еталонної оберемо вибірку 2. Об'єм еталонної вибірки за таблицею 1 становить n=681. Будемо по черзі розглядати попадання спостереження з еталонної вибірки в певну групу як «успіх», а в решту – як «невдачу». Тоді ми матимемо справу з біноміальними розподілами, для теоретичних ймовірностей яких легко побудувати двосторонній асимптотичний довірчий інтервал будь-якої наперед заданої надійності [14].

Наприклад, попадання значення еталонної вибірки в групу 1 будемо вважати «успіхом», а в одну з решти груп – «невдачею». Вибіркова оцінка невідомого стандартного відхилення ймовірності «успіху» для еталонної генеральної сукупності з n=681

$$s_n = \sqrt{\frac{w_{усп}(1-w_{усп})}{n}} \approx 0.017$$

$w_{усп} \approx 190/681 \approx 0{,}28$ (див. таблицю 1).

Тоді права межа довірчого інтервалу надійності 99% становить

$w_{усп} + t_{0,99} \cdot s_n \approx 0{,}32$,

а ліва межа -

$w_{усп} - t_{0,99} \cdot s_n \approx 0{,}23$,

$t_{0,99} \approx 2{,}58$ для двостороннього розподілу Стьюдента з n-1=680 ступенями вільності (в Excel-2010 його можна знайти за допомогою функції «СТЬЮДЕНТ.ОБР.2Х(0,01;680)»).

Тобто, з надійністю 99% теоретична ймовірність того, що довільно обраний елемент еталонної генеральної сукупності належить групі 1, знаходиться в інтервалі (0,23; 0,32).

Так само побудуємо довірчі інтервали надійності 99% для теоретичних ймовірностей, що відповідають іншим групам (таблиця 2).

Таблиця 2. Межі довірчих інтервалів надійності 99% для невідомих теоретичних ймовірностей еталонної генеральної сукупності

Групи	Ліва межа довірчого інтервалу	Права межа довірчого інтервалу
Група 1	0,23	0,32
Група 2	0,21	0,30
Група 3	0,08	0,14
Група 4	0,15	0,22
Група 5	0,14	0,21

χ^2 – відстань між емпіричним та теоретичним розподілом знаходять за формулою [15]

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{r}\frac{(\nu_i^k - np_i)^2}{p_i} \quad (1)$$

Якщо спостережена вибірка належить саме тій генеральній сукупності, теоретичні ймовірності якої розглядаються, то величина (1) має асимптотичний χ^2 – розподіл з r-1 ступенями вільності. Залишається порівняти отриману відстань (1) з критичним значенням, що відповідає певному рівню істотності α. Якщо відстань не перевищує критичного значення (в Excel 2010 його можна знайти за допомогою функції ХИ2.ОБР.ПХ(α;r-1)) приймають рішення про справедливість нульової гіпотези, що вибірковий розподіл співпадає з теоретичним, у протилежному випадку – приймають альтернативну гіпотезу.

Однак, в нашому випадку ймовірності p_i невідомі, проте відомі довірчі інтервали для них (див. табл. 2), побудовані за еталонною вибіркою. Замінимо p_i на такі значення з довірчих відрізків Δ_i (тобто з довірчих інтервалів, разом із їх кінцями), які зроблять значення виразу (1) найменшим з можливих. Тим самим при перевірці гіпотези про узгодженість розподілів ми мінімізуємо помилку 1-го роду. Таким чином, замість відстані (1) будемо розглядати відстань

$$\tilde{\chi}^2 = \min_{p_i\in\Delta_i, i=1,\dots,r}\left(\frac{1}{n}\sum_{i=1}^{r}\frac{(\nu_i^k - np_i)^2}{p_i}\right) \quad (2)$$

Будемо вважати, що вибірки належать одному кластеру, якщо побудована для них відстань (2) до еталонного теоретичного розподілу не більша за критичне значення $\chi^2_{кр}$=*ХИ2.ОБР.ПХ(α;r-1).*

Зауважимо, що довірчі відрізки не повинні містити нульові значення. Якщо це трапилося, тобто значення відносних частот для деяких категорій дуже малі, варто об'єднати ці категорії з іншими.

Якщо формування першого кластеру закінчено, то в якості еталонної обирається вибірка найбільшого об'єму з числа тих, що не попали у перший кластер, та процес кластеризації продовжується, і т.д.

Визначення вибірки найбільшого об'єму в якості еталонної призводить до звуження довірчого інтервалу, тобто збільшення точності оцінювання. Крім того, якщо на кожному кроці створення нового кластеру існує лише одна вибірка з найбільшим об'ємом, то результат кластеризації стає однозначним.

В нашому прикладі на першому кроці вибірки 1, 3, 4 почергово за формулою (2) порівнюються з критичним значенням $\chi^2_{кр}$=*ХИ2.ОБР.ПХ(0,01;2)≈9,21,* де значення ν_i^k беруться з таблиці 1, а довірчі відрізки Δ_i , побудовані за еталонною вибіркою (в нашому випадку - вибіркою 2) з таблиці 2. Знаходження мінімального значення виразу (2) можна здійснити за допомогою інструменту Excel «Пошук розв'язків». Процес знаходження мінімуму виразу (2) спроститься, якщо зауважити, що

$$\min_{p_i\in\Delta_i, i=1,\dots,r}\left(\frac{1}{n}\sum_{i=1}^{r}\frac{(\nu_i^k - np_i)^2}{p_i}\right) = \frac{1}{n}\sum_{i=1}^{r}\min_{p_i\in\Delta_i}\left(\frac{(\nu_i^k - np_i)^2}{p_i}\right)$$

Оцінимо тепер ймовірність помилкової кластеризації, точніше ймовірність того, що при правильності нульової гіпотези про однорідність усіх теоретичних розподілів, з яких узяті спостережені вибірки, знайдеться принаймні пара вибірок, які попадуть у різні кластери.

Нехай всі довірчі інтервали, які ми будували для невідомих ймовірностей еталонного теоретичного розподілу (див., наприклад, табл. 2), мають однакову надійність, рівну γ. Точні значення ймовірностей еталонного теоретичного розподілу нам принципово невідомі. Розглянемо подію *A*, що усі *r* ймовірностей еталонного теоретичного розподілу попадуть у відповідні довірчі інтервали. Оскільки відповідні довірчі інтервали будуються незалежно, то ймовірність $P(A)=\gamma^r$. Нехай подія *B* означає, що за критерієм χ^2 помилково прийняте рішення, що не всі генеральні сукупності мають однакові розподіли, хоча в дійсності всі вони однорідні. Ймовірність цієї події – це

рівень істотності α, на якому перевіряється гіпотеза про однорідність генеральних розподілів, з яких узяті спостережені вибірки. Для визначеності будемо обирати $\alpha=1-\gamma$. Помилка кластеризації відбувається, коли наступає подія $\bar{A}$ або одночасно наступають події A та B. Оскільки ці події несумісні, то ймовірність помилки кластеризації не більша за

$$P(\bar{A})+P(A\cdot B)=P(\bar{A})+P(B/A)P(A)=1-\gamma^r+\alpha\gamma^r.$$

Якщо прийняти $\alpha=1-\gamma$, то ймовірність помилки кластеризації не перевищуватиме величини $1-\gamma^{r+1}$.

Якщо ми хочемо задати певний рівень похибки кластеризації p, то відповідне значення надійності довірчих інтервалів γ, а отже і значення рівня істотності $\alpha=1-\gamma$, що забезпечують похибку кластеризації не вищу за p, можна знайти з рівняння

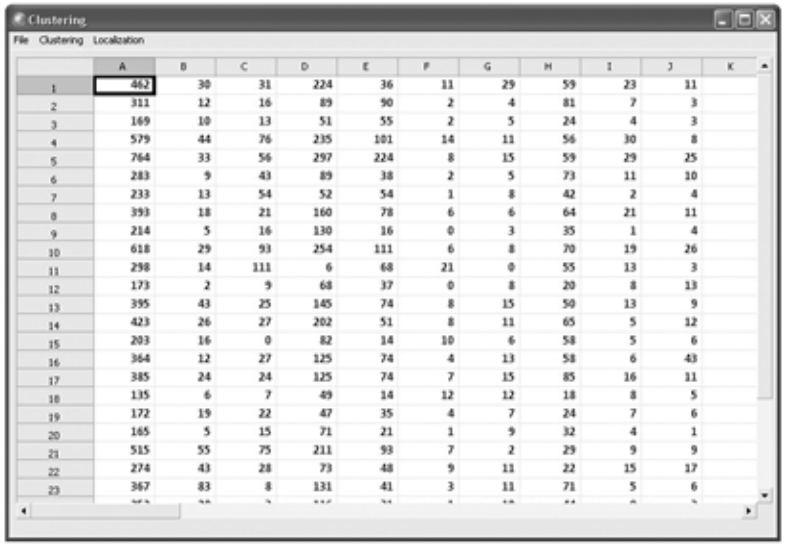
$$1-\gamma^{r+1}=p,$$

звідки

$$\gamma=(1-p)^{\frac{1}{r+1}} \quad (3)$$

Скажімо, в нашому прикладі (таблиця 1) для рівня похибки кластеризації p=0,05 необхідно будувати довірчі інтервали надійності $\gamma=0{,}95^{(1/6)}\approx 0{,}99$, в цьому випадку рівень істотності при перевірці за критерієм χ^2 гіпотези про однорідність генеральних розподілів буде вважатися рівним $\alpha=1-\gamma\approx 0{,}01$.

Результати. Реалізація запропонованого методу здійснена в інтегрованому середовищі розробки програмного забезпечення для Microsoft Windows, Mac OS, iOS і Android на мові Delphi (RAD - Rapid Application Development) Delphi XE6. Бібліотека наявних компонент дозволила реалізувати введення даних, проведення кластеризації та виведення отриманого результату в одному додатку (рис. 1).



	A	B	C	D	E	F	G	H	I	J	K
1	462	30	31	224	36	11	29	59	23	11	
2	311	12	16	89	90	2	4	81	7	3	
3	169	10	13	51	55	2	5	24	4	3	
4	579	44	76	235	101	14	11	56	30	8	
5	764	33	56	297	224	8	15	59	29	25	
6	283	9	43	89	38	2	5	73	11	10	
7	233	13	54	52	54	1	8	42	2	4	
8	393	18	21	160	78	6	6	64	21	11	
9	214	5	16	130	16	0	3	35	1	4	
10	618	29	93	254	111	6	8	70	19	26	
11	298	14	111	6	68	21	0	55	13	3	
12	173	2	9	68	37	0	8	20	8	13	
13	395	43	25	145	74	8	15	50	13	9	
14	423	26	27	202	51	8	11	65	5	12	
15	203	16	0	82	14	10	6	58	5	6	
16	364	12	27	125	74	4	13	58	6	43	
17	385	24	24	125	74	7	15	85	16	11	
18	135	6	7	49	14	12	12	18	8	5	
19	172	19	22	47	35	4	7	24	7	6	
20	165	5	15	71	21	1	9	32	4	1	
21	515	55	75	211	93	7	2	29	9	9	
22	274	43	28	73	48	9	11	22	15	17	
23	367	83	8	131	41	3	11	71	5	6	

Рис 1**.** Додаток в режимі введення даних

Початкове введення даних для кластеризації запропонованим методом здійснюється в самому додатку. Також передбачено експорт підготовлених даних з таблиці Microsoft Excel.

Для налаштування додатку необхідно задати масштаб (коефіцієнт множення), номер еталонної категорії та рівень значущості в інтерактивному режимі.

Результат, отриманий внаслідок кластеризації, представляється у вигляді схеми (рис. 2), де представлено номери отриманих кластерів та номери категорій (рядків), які входять до даного кластера.

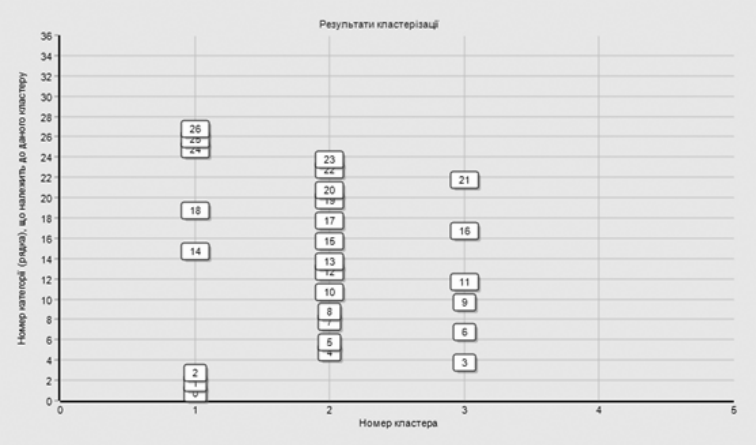


Рис 2. Представлення результатів кластеризації

Запропоноване представлення результатів дозволяє наочно спостерігати розподіл даних по кластерах.

Вихідні дані та результати кластеризації зберігаються в таблиці Microsoft Excel. Також надається можливість збереження результатів кластеризації в графічному форматі.

Для комфортного використання передбачено зміну мови інтерфейсу додатку (англійська та українська).

За офіційними даними аварійності [16] на основі розглянутого вище алгоритму за допомогою розробленої нами в середовищі Delphi програми був проведений кластерний аналіз регіонів України за розподілом дорожньо-транспортних пригод з постраждалими у 2016 році по видах, винуватцях та основних причинах скоєння. Ймовірність похибки кластеризації була обрана рівною 5%. Результати аналізу зведемо у таблиці 3, 4, 5.

Таблиця 3. Кластеризація регіонів за розподілом ДТП з постраждалими у 2016 році по винуватцях

Область або місто	Кластер	Причини ДТП з постраждалими		
		з вини водіїв,	з вини дорослих пішоходів,	з вини дітей,
		к-сть	к-сть	к-сть
Вінницька	1	563	61	22
Волинська	1	480	61	24
Дніпропетровська	1	1926	320	71
Донецька	1	593	68	21
Житомирська	2	345	24	10
Закарпатська	2	201	8	9
Запорізька	1	740	84	18
Івано-Франківська	1	440	89	22
Київська	2	369	17	7
Київ	2	734	44	5
Кіровоградська	2	250	16	10
Луганська	2	232	11	13
Львівська	2	702	46	19
Миколаївська	1	303	32	6
Одеська	2	1073	83	33
Полтавська	1	752	76	19
Рівненська	1	310	40	13

Сумська	1	341	38	8
Тернопільська	1	272	26	12
Харківська	2	443	28	7
Херсонська	1	357	32	16
Хмельницька	2	260	12	9
Черкаська	2	713	52	22
Чернігівська	2	494	56	16
Чернівецька	1	164	4	4

Таблиця 4. Кластеризація регіонів за розподілом ДТП з постраждалими у 2016 році по їх видах

Область або місто	Кластер	Види ДТП					
		Зіткнення,	Перекидання,	Наїзд на ТЗ, що стоїть,	Наїзд на перешкоду,	Наїзд на пішохода,	Наїзд на велосипедиста,
		к-сть	к-сть	к-сть	к-сть	к-сть	к-сть
Вінницька	1	375	72	12	95	382	86
Волинська	4	285	91	17	92	307	97
Дніпропетровська	1	947	136	81	276	880	126
Донецька	1	348	41	25	138	328	58
Житомирська	1	353	67	24	118	323	84
Закарпатська	5	197	29	11	108	176	53
Запорізька	1	454	89	15	126	399	73
Івано-Франківська	5	216	32	4	67	259	60
Київська	2	574	63	32	174	399	91
Київ	1	1027	24	48	191	1077	75
Кіровоградська	1	217	50	13	50	146	32
Луганська	4	131	40	13	53	67	34
Львівська	1	836	125	22	228	717	105
Миколаївська	1	405	66	20	80	324	37
Одеська	1	911	147	68	241	682	88
Полтавська	4	399	97	42	104	315	104
Рівненська	1	298	81	24	87	281	68
Сумська	1	205	58	18	57	203	50
Тернопільська	1	223	34	14	46	190	28
Харківська	3	720	86	42	113	558	58
Херсонська	1	285	63	32	82	234	24
Хмельницька	1	312	47	17	82	272	43
Черкаська	1	359	67	29	125	238	66
Чернігівська	1	253	62	17	82	216	105
Чернівецька	4	115	21	7	45	130	16

Таблиця 5. Кластеризація регіонів за розподілом ДТП з постраждалими у 2016 році по причинах скоєння

Область або місто	Кластер	Керування у нетверезому стані	Перевищення безпечної швидкості	Порушення правил маневрування	Порушення правил проїзду пішохідних	Порушення правил обгону	Виїзд на смугу зустрічного руху	Порушення правил проїзду перехресть	Недодержання дистанції
	к-сть	к-сть	к-сть	к-сть	к-сть	к-сть	к-сть	к-сть	к-сть
Вінницька	3	103	194	151	42	13	47	34	39
Волинська	3	113	177	90	58	14	28	54	23
Дніпропетровська	1	124	616	319	227	26	104	241	267
Донецька	2	78	219	118	35	6	27	132	60
Житомирська	6	93	99	42	31	7	30	38	30
Закарпатська	4	29	97	42	5	4	11	7	20
Запорізька	3	115	186	222	70	9	39	92	86
Івано-Франківська	1	60	210	89	52	14	40	48	43
Київська	3	116	116	77	15	6	15	37	33
Київ	6	78	172	220	111	2	18	148	142
Кіровоградська	1	22	98	53	8	9	17	43	26
Луганська	2	64	100	52	6	2	14	44	15
Львівська	4	98	420	130	25	17	46	46	53
Миколаївська	2	53	114	55	17	11	7	57	21
Одеська	2	134	359	198	65	15	65	170	139
Полтавська	2	107	273	179	61	12	60	110	82
Рівненська	4	45	179	41	13	9	21	23	9
Сумська	1	43	114	64	44	8	21	38	44
Тернопільська	3	48	64	48	37	11	26	37	20
Харківська	4	88	215	95	1	5	17	32	31
Херсонська	2	51	155	58	14	12	20	54	30
Хмельницька	2	54	71	60	15	3	24	15	31
Черкаська	5	116	328	140	42	12	37	106	38
Чернігівська	3	109	279	83	15	10	57	55	58
Чернівецька	4	18	73	69	9	9	21	10	11

Запропонований алгоритм кластерного аналізу даних по кількох факторах (вимірах) забезпечує, на відміну від інших поширених методів, однозначність розбиття на кластери з одночасною оцінкою ймовірності похибки 1-го роду, тобто ймовірності, що однорідні дані будуть віднесені до різних кластерів. Таким чином, його перевагою є той факт, що він дозволяє проводити не тільки розвідувальний аналіз даних, але й робити математично обгрунтовані висновки. Необхідною умовою для його застосування є наявність інформації про кількість спостережень кожного фактору для кожного об'єкту, що кластеризується.

Так, застосування цього методу до аналізу регіонів України щодо аварійності на автомобільному транспорті з постраждалими дозволило з ймовірністю похибки, що не перевищує 5%, згрупувати ці регіони по видах ДТП, винуватцях та причинах скоєння.

Висновки. Запропонований метод кластеризації незалежних наборів багатовимірних даних за допомогою критерію узгодженості $\chi 2$ відрізняється обчислювальною простотою, однозначністю розбиття даних на кластери, що не перетинаються, а також можливістю контролю похибки кластеризації. Цей метод був створений з метою порівняння структурних відмінностей багатовимірних систем різної природи. Кластеризація регіонів України розглянутим методом за розподілом кількості ДТП з постраждалими залежно від різних факторів дозволяє в якості наступного кроку провести аналіз спектру організаційних, технічних, кадрових та інших причин, які призвели до об'єднання регіонів в один та різні кластери, для поширення кращих практик запобігання аварійності на автомобільному транспорті в Україні та напрацювання заходів щодо зниження її загального рівня.

Список бібліографічних посилань

1. Олдендерфер М. С., Блэшфилд Р. К. Кластерный анализ // Факторный, дискриминантный и кластерный анализ: пер. с англ. / под. ред. И. С. Енюкова. — М.: Финансы и статистика, 1989. — 215 с.
2. Гройсман В. Безпека руху на українських дорогах має бути відчутна кожному водію [Електрон. ресурс] / В. Гройсман. – Департамент інформації та комунікацій з громадськістю Секретаріату Кабінету Міністрів України, опубліковано 10 квітня 2018 року. – Режим доступу: https://www.kmu.gov.ua/ua/news/bezpeka-ruhu-na-ukrayinskih-dorogah-maye-buti-vidchutna-kozhnomu-vodiyu-volodimir-grojsman
3. Семь основных причин ДТП [Электрон. ресурс]. – Режим доступа: https://auto.tsn.ua/obzory/7-osnovnyh-prichin-dtp-418849.html
4. В Украине за 1,5 года на дорогах погибло больше людей, чем в АТО [Электрон. ресурс]. – Режим доступа: https://inforesist.org/v-ukraine-za-1-5-goda-na-dorogah-pogiblo-bolshe-lyudey-chem-v-ato/
5. Ханін О. Г. Методологічні особливості застосування критерію узгодженості χ^2 в практичних задачах економіки, соціології та маркетингу / О. Г. Ханін // Економічний аналіз: зб. наук. праць / Тернопільський національний економічний університет. – 2015. – Том 22. – № 1. – С. 67–70.
6. Ханін О. Г. Метод χ^2-кластеризації в задачах маркетингу / О. Г. Ханін // Економічний аналіз: зб. наук. праць / Тернопільський національний економічний університет. – 2016. – Том 26. – № 1. – С. 38–42.
7. Черезов Д.С., Тюкачев Н.А. Обзор основных методов классификации и кластеризации данных // Вестник Воронежского государственного университета, - Серия: Системный анализ и информационные технологи, - №2, 2009, с.25-29
8. Tryon, R.C. (1939) Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers, Ann Arbor. - 122 p.
9. Trebuňa P., Halčinová J. Mathematical Tools of Cluster Analysis // Applied Mathematics, 2013, 4, 814-816.
10. Нейский И. М. Классификация и сравнение методов кластеризации [Електронний ресурс] / И. М. Нейский.- Режим доступу: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf.
11. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. 3,- 69 p.
12. Тьюки Дж. Анализ результатов наблюдения. Разведочный анализ. М.: Мир,- 1981, - 696 с.
13. Кластерный анализ [Электрон. ресурс]. – Режим доступа: http://ieee.tpu.ru/system/cluster.html
14. Сигел Э. Практическая бизнес-статистика / Э. Сигел – М. : Вильямс, 2002. – 1056 с.
15. Крамер Г. Математические методы статистики / Г. Крамер – М. : Мир, 1976. – 648 с.
16. Статистика аварійності в Україні за 12 місяців 2016 року [Електрон. ресурс]. – Режим доступу: http://www.sai.gov.ua/ua/ua/static/21.htm (дата звернення: 05.06.2017)