РОЗДІЛ 8
СТАТИСТИКА

УДК 311:314

Хвалинська В. В.
Інститут демографії та соціальних досліджень імені М.В. Птухи
Національної академії наук України

КЛАСТЕРИЗАЦІЯ РЕГІОНІВ УКРАЇНИ ЗА ДЕМОГРАФІЧНИМИ ПОКАЗНИКАМИ: ГРУПУВАННЯ НА ОСНОВІ АНСАМБЛЮ АЛГОРИТМІВ

У статті розглядається метод ансамблевого кластерного аналізу, заснований на знаходженні узгодженої матриці подібності об'єктів. Пропонується ансамбль алгоритмів кластеризації, який складається з різних агломеративних ієрархічних методів, що відрізняються мірою об'єднання кластерів (обчисленням відстані між кластерами). Наводяться результати кластеризації регіонів України за демографічними показниками з використанням запропонованого ансамблю алгоритмів.
Ключові слова: кластерний аналіз, ансамблевий підхід, відстані між кластерами, демографічні дослідження.

Постановка проблеми. В останні роки у соціальних дослідженнях усе більше уваги приділяється методам багатовимірного статистичного аналізу, які відзначаються широким спектром можливостей під час моделювання і структуризації масових процесів та явищ, що характеризуються великою кількістю ознак. Зокрема, зростає популярність кластерного аналізу, основною метою якого є виділення порівняно невеликої кількості груп об'єктів, які всередині групи схожі між собою і відрізняються у різних групах. Автори використовують різні методологічні підходи до проведення класифікації за допомогою методів кластерного аналізу. Відомо, що алгоритми кластерного аналізу не є універсальними, кожний має специфічну сферу використання, переваги і недоліки. Актуальною у кластерному аналізі залишається проблема якості й стійкості розбиття. Останнім часом для вирішення цієї проблеми активно застосовується ансамблевий підхід. Суть його полягає у тому, що для знаходження кластерів застосовується не один алгоритм, а набір різноманітних алгоритмів, а потім на їх основі будується підсумкове розбиття.

Аналіз останніх досліджень і публікацій. У роботі [1, с. 14–18] описується ансамблевий підхід до розв'язання задачі кластер-аналізу. Авторами розглядаються різноманітні принципи вибору найкращої узгоджувальної функції. У роботі [2, с. 24–25] для групування результатів та визначення узагальненого розв'язку виділено найбільш популярні підходи: прямий підхід, або перевизначення кластерів, графовий, матричний та ймовірнісний. Ідея побудови колективних розв'язків, заснованих на комбінації простих алгоритмів, активно використовується у різних прикладних сферах: медицині, гідрохімії, під час аналізу гіперспектральних зображень [2–4] та ін.

Аналіз джерел свідчить про актуальність даного напряму досліджень, оскільки застосування ансамблевого підходу знижує залежність результатів групування від вибору параметрів алгоритму, підвищує якість та стійкість результатів, зменшує простір ознак, кластеризує різнотипні дані та вирішує багато інших завдань.

Виділення невирішених раніше частин загальної проблеми. Нині розроблена велика кількість методів і алгоритмів кластерного аналізу [5, с. 48–52]. У класичних алгоритмах розв'язання задач кластер-аналізу результати групування можуть суттєво змінюватися залежно від вибору системи ознак, мір близькості, вибору початкових умов, порядку об'єктів, параметрів роботи алгоритму тощо. Підвищити ефективність кластерного аналізу можна, застосувавши ансамблевий підхід, який полягає у побудові множини кластеризацій на основі різноманітних алгоритмів або одного алгоритму з різними параметрами і підсумковій кластеризації на їх основі. Таким чином, розробляючи ансамбль із різним набором алгоритмів з їх перевагами і особливостями, можна розробити найбільш привабливу схему кластеризації для вирішення конкретного завдання визначеної предметної сфери.

Мета статті полягає у розробленні і застосуванні ансамблю алгоритмів для кластеризації регіонів України за демографічними показниками на основі ієрархічних агломеративних методів із використанням узгодженої матриці подібності, а також аналізі отриманих результатів.

Виклад основного матеріалу дослідження. Існує декілька основних способів побудови колективних розв'язків кластерного аналізу [1, с. 14–18]. Одним із найбільш поширених є підхід, який базується на обчисленні узгодженої матриці подібності об'єктів, яка визначає, як часто пара об'єктів потрапляє до складу одного кластеру в різних варіантах розбиття.

Групування на основі узгодженої матриці подібності. Нехай використовуються L методів кластеризації. На першому етапі будується набір групувань (розв'язків кластеризації) $G = \{G^{(1)}, G^{(2)},..., G^{(L)}\}$, де $G^{(i)}$ – i-й варіант групування, який містить $K^{(i)}$ кластерів. Вводиться для i-го групування бінарна матриця подібності $S^{(i)} = \{S^{(i)}(j, m)\}$ розмірністю $N \times N$ (N – число об'єктів, що кластеризуються) так: $S^{(i)}(j, m) = 1$, якщо об'єкти $o^{(j)}$ та $o^{(m)}$ належать одному кластеру і $S^{(i)}(j, m) = 0$, якщо не належать, де $j, m = 1, 2, ..., N$; $i = 1, 2, ..., L$.

На другому етапі формується узгоджена матриця подібності $S = \{S(j, m)\}$:

$$S(j,m) = \frac{1}{L}\sum_{i=1}^{L} S^{(i)}(j,m),$$

де $j, m = 1, 2, ..., N$. Величина $S(j, m)$ дорівнює частості класифікації об'єктів $o^{(j)}$ та $o^{(m)}$ до однієї

групи в наборі групувань G. Близьке до одиниці значення величини $S(j, m)$ означає, що дані об'єкти мають великий шанс потрапити до однієї групи. Близьке до нуля значення цієї величини говорить про те, що шанс потрапити до однієї групи у цих об'єктів незначний.

На третьому етапі узгоджена матриця подібності використовується для знаходження підсумкового варіанту групування шляхом застосування методів кластеризації, які як вхідну інформацію використовують відстані між об'єктами (наприклад, ієрархічні методи).

Процедура утворення ансамблю алгоритмів. Для вирішення поставленого завдання на першому етапі в ансамблі алгоритмів кластеризації використовувалися ієрархічні агломеративні методи, принцип роботи яких полягає у тому, що початковим є розбиття на n одноелементних кластерів, які послідовно об'єднуються в один клас: спочатку у групи об'єднуються найбільш близькі об'єкти, а потім – усе більш віддалені. За їх допомогою можна здійснити повний і тонкий аналіз структури досліджуваної сукупності, також варто відзначити зручність використання та можливість наочно представити результати кластеризації. Ці методи групування достатньо повно представлені в різних пакетах прикладних програм для статистичного аналізу даних. У цьому дослідженні було використано SPSS [6].

На третьому етапі для формування найкращого розбиття за узгодженою матрицею вибрано метод Варда. Цей метод спрямований на об'єднання близько розміщених кластерів і «прагне» створювати кластери малого розміру, що в даному разі вирішує поставлене завдання – побудова найкращого розбиття за узгодженою матрицею [7, с. 477–479].

Процедура утворення ансамблю алгоритмів представлена на рис. 1.

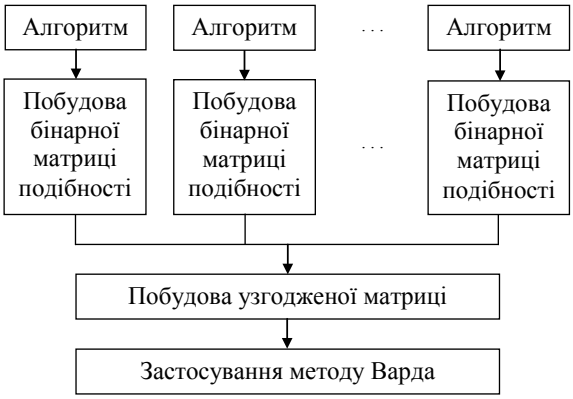

Рис. 1. Ансамбль алгоритмів кластеризації

Джерело: розроблено автором

Кластеризація демографічних показників на основі запропонованого ансамблю алгоритмів. Інформація про об'єкти (регіони України) представлена у вигляді двовимірної таблиці розмірністю $N \times n$: число об'єктів N=25 – Автономна республіка Крим та 24 області (міста Київ та Севастополь не є характерними для всієї сукупності об'єктів), кожен з яких характеризується набором n=9 показників (ознак). Вибір показників здійснювався так, щоб охопити такі демографічні ознаки і характеристики, які дадуть змогу поглибленого вивчення регіональних особливостей демографічної ситуації в Україні. Групування регіонів у кластери відбувалося за такими показниками: сумарний коефіцієнт народжуваності,

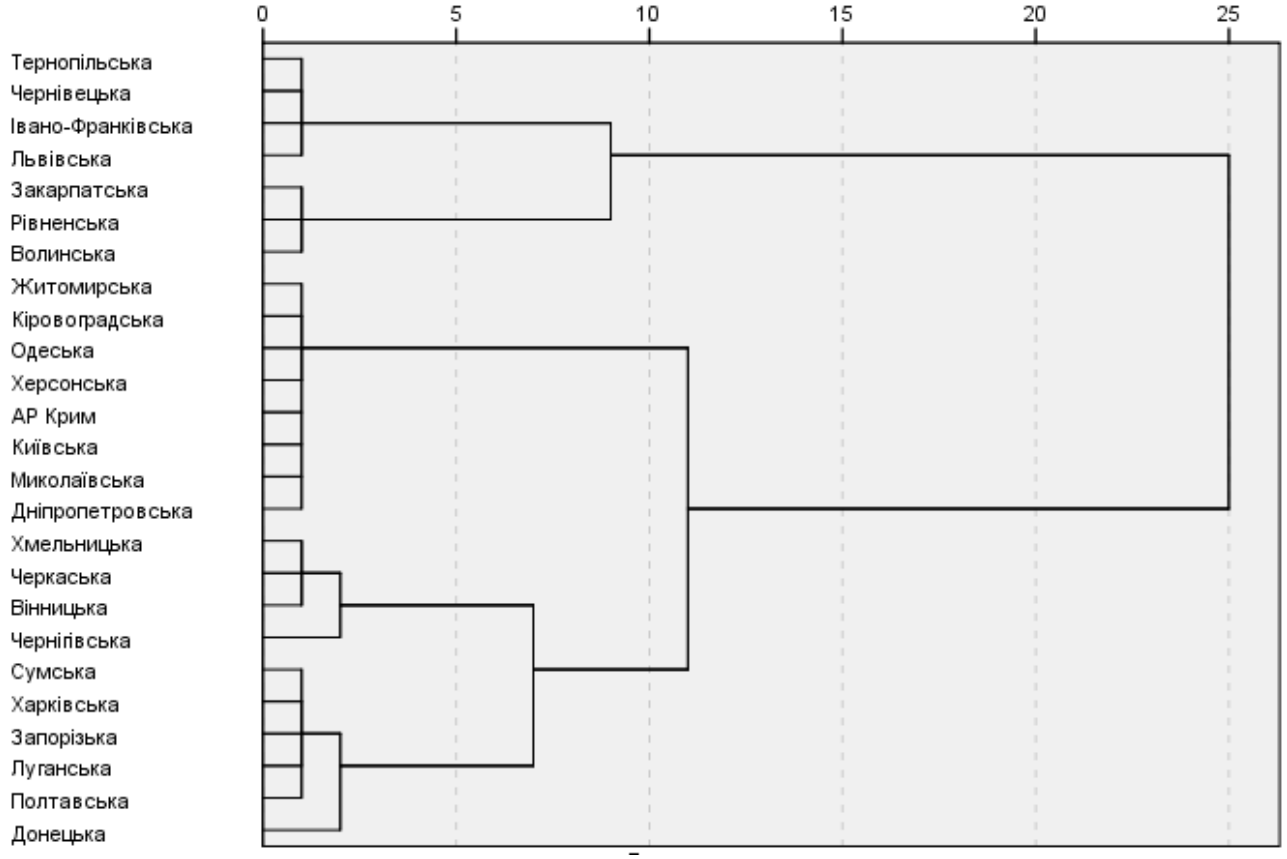

Рис. 2. Дендрограма методу Варда для підсумкового розбиття

Джерело: фрагмент результату роботи процедури кластеризації у SPSS (знімок екрану)

середня очікувана тривалість життя чоловіків і жінок при народженні, середній вік населення, питома вага осіб віком 60 років і старше та дітей віком до 16 років, демографічне навантаження на населення у віці 16–59 років, щільність наявного населення, площа територій. Джерелами статистичних даних є демографічний щорічник «Населення України» за 2012–2013 рр., офіційний сайт Державної служби статистики України [8]. Дослідження проводилося за даними 2013 р. (останній рік, коли дані по всіх регіонах є повними).

На першому етапі був отриманий набір із п'яти групувань (розв'язків кластеризації), здійснений стандартними ієрархічними методами: міжгрупового зв'язку, внутрішньогрупового зв'язку, «далекого сусіда», центроїдним та методом Варда [7, с. 474–486] .

На заключному етапі в рамках ансамблевого підходу для отримання узгодженого розбиття використали ієрархічний метод Варда, дендрограма якого представлена на рис. 2. На ній видно п'ятикластерне розбиття.

Слід зауважити, що на дендрограмі виділяються дві області – Чернігівська та Донецька, які останніми приєднуються до «своїх» кластерів. Це потребує особливої уваги під час опису та аналізу результатів кластеризації.

Аналіз результатів кластеризації регіонів України за демографічними показниками. У результаті застосування запропонованого підходу отримано остаточне розбиття, яке складається із п'яти кластерів. У табл. 1 наведено склад кластерів та показники, за якими проводилося групування.

Ознаками якісної кластеризації є те, що відмінності між утвореними групами повинні бути очевидними, а всередині групи об'єкти повинні бути максимально схожими. Аналіз табл. 1 показує, що всі кластери достатньо чітко розділені за кожною з ознак, а об'єкти одного кластера схожі між собою. Кожний кластер визначається специфічним поєднанням показників, тобто відображає певну демографічну ситуацію і за своїм демографічним змістом відрізняється від інших.

Перший кластер характерний тим, що має найвищі сумарні коефіцієнти народжуваності і найменший середній вік населення, близько 37 років. У другому кластері середня очікувана тривалість

Таблиця 1

Результати кластерного аналізу

	Сумарний коефіцієнт народжуваності, *осіб на одну жінку*	**Середня очікувана тривалість життя при народженні,** *чоловіки*	**Середня очікувана тривалість життя при народженні,** *жінки*	**Середній вік населення,** *років*	**Частка населення у віці 0-15 років у загальній кількості населення,** *%*	**Частка населення у віці 60 років і старше у загальній кількості населення,** *%*	**Демографічне навантаження на населення у віці 16–59 років,** *на 1000 осіб*	**Щільність населення,** *осіб на 1 км²*	**Територія,** *тис. км²*
									1 кластер
Волинська	1,856	66,04	76,92	37,5	20,2	17,7	612	51,6	20,1
Закарпатська	1,926	67,14	74,75	36,6	20,3	16,2	575	98,4	12,8
Рівненська	1,998	66,25	76,53	36,8	21,0	16,9	611	57,7	20,1
									2 кластер
Івано-Франківська	1,603	68,41	77,76	38,6	17,9	18,9	582	99,2	13,9
Львівська	1,552	68,37	78,07	39,2	16,8	19,4	567	116,4	21,8
Тернопільська	1,475	68,69	78,39	39,8	16,8	20,7	600	77,9	13,8
Чернівецька	1,630	68,61	77,63	38,5	17,8	19,1	585	112,1	8,1
									3 кластер
Вінницька	1,530	66,89	76,81	41,0	16,0	23,0	640	61,4	26,5
Хмельницька	1,606	66,73	76,93	40,9	16,1	22,3	624	63,7	20,6
Черкаська	1,378	66,92	76,59	42,0	14,4	24,1	625	60,7	20,9
Чернігівська	1,374	64,66	76,12	43,0	13,8	25,4	646	33,8	31,9
									4 кластер
АР Крим	1,666	66,61	76,28	40,4	15,8	21,1	585	75,3	26,1
Дніпропетровська	1,514	64,91	75,26	40,9	15,0	21,9	585	103,6	31,9
Житомирська	1,680	64,00	75,11	40,2	16,9	21,7	629	42,5	29,8
Київська	1,640	65,24	75,44	40,3	15,9	21,1	588	61,2	28,1
Кіровоградська	1,573	64,80	74,80	41,4	15,3	23,4	631	40,5	24,6
Миколаївська	1,545	65,00	75,06	40,3	15,7	21,0	578	47,7	24,6
Одеська	1,652	65,93	74,67	39,5	16,4	20,3	579	71,9	33,3
Херсонська	1,599	64,87	75,03	40,1	16,1	20,7	582	37,9	28,5
									5 кластер
Донецька	1,320	64,85	75,53	42,3	13,4	23,6	589	165,0	26,5
Запорізька	1,432	66,46	76,49	41,5	14,4	22,4	582	65,7	27,2
Луганська	1,295	65,62	75,82	42,3	13,1	23,1	568	84,6	26,7
Полтавська	1,402	65,91	76,19	41,8	14,2	23,4	601	51,1	28,8
Сумська	1,303	65,74	76,22	42,2	13,6	23,3	586	48,0	23,8
Харківська	1,320	67,34	76,61	41,3	13,4	22,0	547	87,4	31,4
Україна	**1,506**	**66,34**	**76,22**	**40,5**	**15,5**	**21,4**	**585**	**75,5**	**603,5**

Джерело: складено за даними [8]

Таблиця 2

Загальні характеристики населення і територій кластерів

Показники	Номер кластеру				
	1	2	3	4	5
Територія кластера, *тис. км²*	53,0	57,6	99,9	226,9	164,4
Частка території у загальній площі країни, %*	8,8	9,5	16,6	37,6	27,2
Чисельність наявного населення, *тис. осіб*	3451,2	5907,0	5287,7	13906,0	13772,7
Питома вага чисельності населення кластеру у загальній кількості населення, %*	7,6	13,0	11,6	30,5	30,2
Щільність населення, *осіб/км²*	65,1	102,6	52,9	61,3	83,8
Частка міського населення, %	45,2	50,9	55,7	68,2	81,2

* Сума рядка не дорівнює 100% (м. Київ, м. Севастополь не включалися до досліджуваної сукупності).

Джерело: розраховано за даними [8]

життя найвища по Україні. Середній вік населення менший за загальнодержавний показник (40,5 років), сумарні коефіцієнти народжуваності вищі за загальнодержавний, який становить 1,506. Третій кластер характерний тим, що має найбільше демографічне навантаження за високої частки осіб віком 60 років і старше частки дітей віком до 16 років на загальнодержавному рівні. У четвертому кластері у цілому показники наближені до загальнодержавних. Варто зазначити, що сумарні коефіцієнти народжуваності вищі за загальнодержавні, а середня очікувана тривалість життя при народженні найнижча в Україні. У п'ятому кластері найнижчі сумарні коефіцієнти народжуваності і найбільший середній вік населення, близько 14% населення – діти віком 0–15 років, 23% – покоління віком 60 років і старше. У даних областях найскладніша демографічна ситуація.

У табл. 2 наведено загальні характеристики населення і територій кластерів на початок 2013 р.

Перший та другий кластери невеликі за територією, але суттєво відрізняються щільністю населення. Четвертий кластер найбільший за територією і займає 37,6% площі країни. Приблизно рівні за чисельністю населення четвертий і п'ятий кластери, де проживає близько 60% населення України (рис. 3).

Рис. 3. Питома вага чисельності населення кластеру у загальній кількості населення, %

Джерело: побудовано за табл. 2

Сільське населення переважає у першому кластері, близько 20% якого діти. Четвертий, п'ятий кластери характерні тим, що переважна більшість населення проживає там у міських поселеннях (рис. 4).

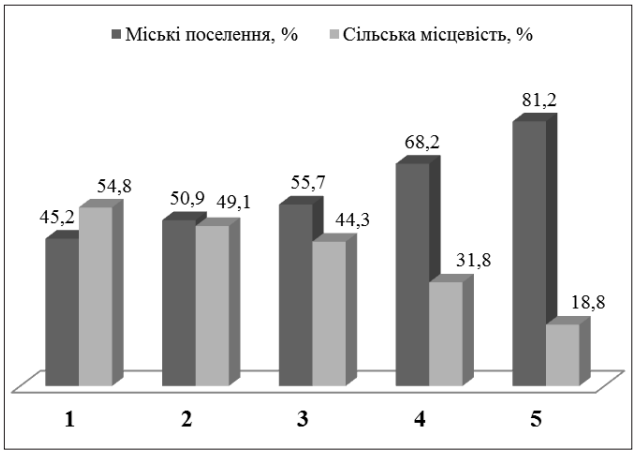

Рис. 4. Частка населення у кластері за типом поселень

Джерело: побудовано за табл. 2

Слід зауважити, що для поглибленого аналізу групувань можна використовувати характеристики кластерів за ознаками, які не були задіяні під час побудови кластерів.

Висновки. У статті розглянуто ансамблевий підхід до розв'язання задачі кластерного аналізу. Запропоновано застосування колективного розв'язку, що базується на побудові узгодженої матриці подібності з використанням ієрархічних алгоритмів із різними мірами об'єднання кластерів. Наведено та проаналізовано результати практичного застосування розробленого ансамблю алгоритмів для кластеризації регіонів України за демографічними показниками.

Перспективами подальших досліджень є вдосконалення запропонованого підходу для підвищення якості отриманих результатів, проведення більш детального аналізу отриманих результатів методами статистичної обробки даних.

Методи побудови оптимальних колективних розв'язків, розроблені з використанням запропонованої моделі, передбачається у подальшому використовувати у соціально-демографічних, соціально-економічних дослідженнях. Суттєвим є те, що цей метод може використовуватися також в інших предметних сферах.

Список використаних джерел:
1. Бериков В.С., Лбов Г.С. Современные тенденции в кластерном анализе. Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». 2008. 26 с.
2. Сидорова М.Г. Застосування ансамблів алгоритмів для підвищення стійкості результатів кластеризації. Актуальні проблеми автоматизації та інформаційних технологій. 2013. Т. 17. С. 22–29.
3. Байбуз О.Г., Сидорова М.Г. Інформаційна технологія нечіткої кластеризації багатовимірних часових рядів на прикладі гідрохімічного моніторингу річки Самара. Науковий вісник Національного гірничого університету. 2014. № 5. С. 114–122.
4. Бериков В.Б. Классификация данных с применением коллектива алгоритмов кластерного анализа. Знания – Онтологии – Теории (ЗОНТ–2015). 2015. Т. 1. С. 29 –38.
5. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.
6. Бюлль А., Цефель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей; пер. с нем. СПб.: ДиаСофтЮП, 2005. 608 с.
7. Многомерный статистический анализ в экономике / Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шеффер. М.: ЮНИТИ-ДАНА, 1999. 598 с.
8. Офіційний сайт Державної служби статистики України. URL: http://www.ukrstat.gov.ua.

Хвалынская В. В.
Институт демографии и социальных исследований имени М. В. Птухи
Национальной академии наук Украины

КЛАСТЕРИЗАЦИЯ РЕГИОНОВ УКРАИНЫ ПО ДЕМОГРАФИЧЕСКИМ ПОКАЗАТЕЛЯМ: ГРУППИРОВКА НА ОСНОВЕ АНСАМБЛЯ АЛГОРИТМОВ

Резюме
В статье рассматривается метод ансамблевого кластерного анализа, основанный на нахождении согласованной матрицы подобия объектов. Предлагается ансамбль алгоритмов кластеризации состоящий из различных агломеративных иерархических методов, которые отличаются мерой объединения кластеров (вычислением расстояния между кластерами). Приводятся результаты кластеризации регионов Украины по демографическим показателям с использованием предложенного ансамбля алгоритмов.
Ключевые слова: кластерный анализ, ансамблевый подход, расстояния между кластерами, демографические исследования.

Khvalynska V. V.
Ptoukha Institute for Demography and Social Studies
of the National Academy of Sciences of Ukraine

CLUSTERING OF THE REGIONS OF UKRAINE BY DEMOGRAPHIC INDICATORS: GROUPING BY THE ENSEMBLE OF ALGORITHMS

Summary
A collective approach to cluster analysis is considered in the paper. The ensemble of algorithms of clustering, consisting of different agglomerative hierarchical methods, differing in the measure of cluster combination (by calculation of distance between clusters) is proposed. The results of clustering of regions of Ukraine according to demographic indicators using the proposed ensemble of algorithms are presented.
Key words: cluster analysis, ensemble approach, distances between clusters, demographic studies.