УДК314:519.237.8

Світлана Бегун – кандидат економічних наук, доцент, доцент кафедри обліку та аудиту Східноєвропейського національного університету імені Лесі Українки

Застосування кластерного аналізу для дослідження демографічної ситуації в регіоні

У статті розглянуто кластеризацію регіонів Волинської області у 2014 р. на основі загальних коефіцієнтів народжуваності та смертності, коефіцієнта фемінізації, кількості шлюбів і загального коефіцієнта міграційного приросту. Показано аналітичні та графічні можливості застосування програмного пакета STATISTICA 6.1 для проведення багатомірного групування при аналізі демографічних процесів окремого регіону. Побудовано дендрограму регіонів Волинської області за показниками відтворення населення у 2014 р. Розглянуто основні описові статистики кожного сформованого кластера.

Ключові слова: кластерний аналіз, демографічна ситуація, відтворення населення, кластеризація, статистичні методи, демографічні процеси, кластери, таксономія, розпізнавання образів.

Постановка наукової проблеми та її значення. Демографічна ситуація, що склалася в Україні за роки незалежності, характеризується скороченням чисельності населення, зростанням смертності, зменшенням народжуваності, значними міграційними процесами, постарінням населення. Усі ці негативні процеси відбуваються під час складних кризових явищ в економіці країни, що сприяють поглибленню демографічної кризи. Протягом 2010–2014 рр. відбуються певні позитивні зрушення щодо покращення демографічної ситуації, особливо помітні в областях Західного регіону країни, серед яких особливе місце займає Волинська область.

Аналіз досліджень цієї проблеми. Проблеми аналізу демографічних процесів в Україні розкрито в роботах Е. Лібанової, І. Курило, С. Пирожкова, С. Аксьонової й ін., застосування методів баговимірного аналізу – О. Гладун, В. Хвалинська, Ю. Феленчак та ін. Питанню кластеризації демографічних процесів населення Волині як області Західного регіону на сьогодні приділено недостатньо уваги.

Формулювання мети та завдань статті. Мета дослідження – розглянути можливості застосування методів кластеризації багатовимірного розвідувального аналізу даних у системі STATISTICA (StatSoft) під час аналізу демографічної ситуації у Волинській області.

Виклад основного матеріалу й обґрунтування отриманих результатів дослідження. Групування та класифікація – це статистичні методи розподілу однорідних і неоднорідних сукупностей на певні групи за істотними ознаками, які знайшли широке застосування в біології, психології, соціології, економіці, менеджменті, демографії тощо. Багатовимірне групування може здійснюватися на основі кластерного аналізу. Кластеризація дає змогу досліджувати великий обсяг інформації, який стосується значної кількості різноманітних ознак, що характеризують сукупність об'єктів, і стискати цю інформацію до зручних, наочних розмірів [6].

Кластерний аналіз (таксономія, розпізнавання образів) складається з різних методів класицікації, основна мета яких – розподіл сукупності об'єктів на невелику (відому чи ні) кількість груп, класів однорідних, схожих об'єктів. Ці групи мають бути сформовані таким чином, щоб об'єкти, які містяться в одному класі, перебували недалеко одна від одного. Такі класи називаються кластерами (таксонами, образами). *Cluster* (англ.) – гроно, пучок, скупчення, група елементів, які мають будь-яку загальну властивість. *Taxon* (англ.) – систематизована група певної категорії [1, с. 489].

На сьогодні існує велика кількість методів кластеризації, в основі яких – різні підходи до виділення кластерів, вибір конкретного методу залежить від практичного застосування отриманого результату.

Кластерний аналіз – достатньо трудомісткий метод статистичного дослідження, тому краще проводити його за допомогою різноманітних програмних продуктів. Система STATISTICA (StatSoft) у середовищі Windows містить у собі всі відомі методи статистичного аналізу даних, що дає змогу зробити процес дослідження більш ефективним і простим [2, с. 41].

У спеціалізованому програмному пакеті STATISTICA 6.1, що використано для проведення класифікації регіонів Волинської області за основними показниками відтворення населення, представлені такі три методи:

- ієрархічна класифікація характеризується побудовою ієрархічної, або деревоподібної, структури. Вона ґрунтується на послідовній кластеризації, основний зміст якої полягає в тому, що спочатку кожен об'єкт є окремим кластером, на наступному кроці найбільш подібні об'єкти об'єднуються в окремий кластер, у подальшому це триває до тих пір, поки всі об'єкти не утворять один кластер;

- кластеризація методом k-середніх – ітеративний метод кластеризації. Був запропонований у 1967 р. Дж. Мак-Куіном [3, с. 20]. Його суть полягає в тому, що процес класифікації починається із задавання певних умов кластерного аналізу, зокрема кількості кластерів, порога завершення процесу класифікації тощо;
- двохвхідне об'єднання передбачає одночасну кластеризацію за змінними (стовпцями) і за результатами спостережень (рядками), проводиться в тих випадках, коли очікується, що одночасна кластеризація за змінними та спостереженнями дасть можливість отримати осмислені результати. На практиці застосовується достатньо рідко.

Для аналізу процесів відтворення населення Волинської області сформовано сукупність із 20 об'єктів, серед них – чотири найбільші міста обласного підпорядкування (Луцьк (1), Володимир-Волинський (2), Ковель (3), Нововолинськ (4)) і 16 районів області, зокрема Володимир-Волинський (5), Горохівський (6), Іваничівський (7), Камінь-Каширський (8), Ківерцівський (9), Ковельський (10), Локачинський (11), Луцький (12), Любешівський (13), Любомльський (14), Маневицький (15), Ратнівський (16), Рожищенський (17), Старовижівський (18), Турійський (19), Шацький (20).

Для виявлення територіальної подібності населення між районами та містами області сформовано сукупність змінних для характеристики процесів відтворення населення Волинської області, яка складається з п'яти показників за 2014 р.: загальний коефіцієнт народжуваності, *‰;* загальний коефіцієнт смертності, *‰;* коефіцієнт фемінізації (відношення чисельності жінок до кількості чоловіків), *‰;* кількість шлюбів, *одиниць;* загальний коефіцієнт міграційного приросту, *‰.*

Вихідні дані для аналізу наведено в табл. 1.

Таблиця 1

Основні показники відтворення населення Волинської області у 2014 р.

	1 Загальний коефіцієнт народжуваності, ‰	2 Загальний коефіцієнт смертності, ‰	3 Коефіцієнт фемінізації,‰	4 Кількість шлюбів, одиниць	5 Загальний коефіцієнт міграційного приросту, ‰
м.Луцьк	12,5	10	1218	1892	2,3
м.Володимир-Волинськи	10,5	10,3	1145	345	2,6
м.Ковель	13,9	10,1	1142	572	0,7
м.Нововолинськ	10,8	12,4	1173	402	0,9
Володимир-Волинський	14	18,1	1105	125	1,3
Горохівський	12,8	16,7	1134	292	-2,7
Іваничівський	11,8	14,7	1099	148	1,2
Камінь-Каширський	17,9	12,7	1055	392	0
Ківерцівський	16,4	14,9	1078	381	-1,1
Ковельський	16,4	16,4	1135	162	-1
Локачинський	12,8	16,5	1101	119	-2
Луцький	15,1	12,1	1095	360	7,7
Любешівський	16,5	12,8	1018	217	-1,7
Любомльський	14,7	16	1084	253	1
Маневицький	14,8	13,8	1059	381	-3,1
Ратнівський	16,1	13,1	1071	335	-1,2
Рожищенський	14	16,2	1117	267	-4
Старовижівський	15,8	17,1	1060	173	1
Турійський	14,4	17,3	1088	160	3,8
Шацький	13	13,9	1066	96	0,8

Сформовано на основі джерела [4].

Мета кластерного аналізу – розподіл територіальних одиниць області на класи (кластери), кожен із яких має певний рівень вибраних демографічних показників. Райони та міста, що потрапляють в один

кластер, характеризуються максимально подібною демографічною ситуацією, тоді як райони та міста з різних кластерів повинні максимально відрізнятися.

У зв'язку з тим, що відібрані змінні мають різні одиниці виміру, виникає необхідність здійснити їх стандартизацію, тобто замінити вихідні дані на нормовані, які дають змогу усунути можливий вплив одиниць виміру. Стандартизацію проводимо автоматично, для цього вибираємо в меню програмного пакета STATISTICA 6.1 «Данные-Стандартизировать». Значення змінних після цього матимуть середнє 0 і стандартне відхилення 1. Отримані стандартизовані (нормовані) дані представлено в табл. 2.

Таблиця 2

Стандартизовані значення показників відтворення населення Волинської області в 2014 році

	1 Загальний коефіцієнт народжуваності	2 Загальний коефіцієнт смертності	3 Коефіцієнт фемінізації	4 Кількість шлюбів	5 Загальний коефіцієнт міграційного приросту
м.Луцьк	-0,8566	-1,6895	2,5067	4,0197	0,7432
м.Володимир-Волинський	-1,8585	-1,5704	0,9272	-0,0225	0,8561
м.Ковель	-0,1553	-1,6498	0,8623	0,5707	0,1411
м.Нововолинськ	-1,7082	-0,7366	1,5330	0,1265	0,2164
Володимир-Волинський	-0,1052	1,5267	0,0617	-0,5973	0,3669
Горохівський	-0,7063	0,9708	0,6892	-0,1610	-1,1383
Іваничівський	-1,2072	0,1767	-0,0682	-0,5372	0,3293
Камінь-Каширський	1,8484	-0,6174	-1,0202	0,1003	-0,1223
Ківерцівський	1,0970	0,2561	-0,5225	0,0716	-0,5362
Ковельський	1,0970	0,8517	0,7108	-0,5006	-0,4986
Локачинський	-0,7063	0,8914	-0,0249	-0,6130	-0,8749
Луцький	0,4458	-0,8557	-0,1547	0,0167	2,7752
Любешівський	1,1471	-0,5777	-1,8208	-0,3569	-0,7620
Любомльський	0,2455	0,6929	-0,3927	-0,2629	0,2540
Маневицький	0,2955	-0,1807	-0,9337	0,0716	-1,2888
Ратнівський	0,9468	-0,4586	-0,6740	-0,0486	-0,5739
Рожищенський	-0,1052	0,7723	0,3213	-0,2263	-1,6275
Старовижівський	0,7965	1,1297	-0,9120	-0,4719	0,2540
Турійський	0,0952	1,2091	-0,3062	-0,5059	1,3077
Шацький	-0,6061	-0,1410	-0,7822	-0,6731	0,1787

Для проведення кластерного аналізу найчастіше на практиці застосовується ієрархічна класифікація. Для цього в меню пакета STATISTICA 6.1 вибираємо «Анализ-многомерный разведочный анализ-кластерный анализ-иерархическая классификация» [6].

Велике значення при кластеризації має вибір адекватної міри близькості між досліджуваними об'єктами. У пакеті STATISTICA 6.1 можна використати такі міри: евклідова відстань, квадрат евклідової відстані, манхетенська відстань, відстань Чебишева, відсоток незгоди, 1-r Пірсона тощо. У [5, с. 17] ще наведено відстань Махаланобіса, міру l1-норму, міру супрерум-норму, міру lp-норму. Але найбільш популярною в алгоритмах кластерного аналізу є міра «евклідова відстань», яка найчастіше застосовується в прикладних дослідженнях [5, с. 16].

Після вибору міри близькості потрібно вибрати правило об'єднання об'єктів у кластери. У пакеті STATISTICA 6.1 можна застосувати такі методи: одиночного зв'язку, що вимірюється за принципом «найближчого сусіда»; повного зв'язку, що вимірюється за принципом «найдальшого су-

сіда», зваженого й незваженого попарного середнього, що вимірюється за принципом «середнього зв'язку»; зваженого та незваженого центроїдного методу, що вимірюється за «центрами ваги» [1, с. 495–496] Варда.

Для проведення кластеризації районів і міст Волинської області вибрано один із найбільш ефективних методів – метод Варда (Ward'smethod), який відрізняється від усіх інших застосовуванням методів дисперсійного аналізу для оцінки відстаней між кластерами. Метод Варда грунтується на внутрішньогруповій сумі квадратів відхилень, яка є сумою квадратів відстаней між кожним об'єктом і середнім значенням у кластері, де розміщений цей об'єкт. При цьому на кожному кроці об'єднуються такі два кластери, які спричиняють найменше зростання внутрішньогрупової суми квадратів. Цей метод спрямований на об'єднання найближчих кластерів [5, с. 30–31].

Перевагою ієрархічних методів класифікації є їх наочність. Результати кластеризації представляються у вигляді дендрограм (у перекладі з грецької dendron означає дерево). Дендрограма наочно зображує близькість окремих об'єктів, кластерів і в графічному вигляді показує послідовність їх об'єднання. Дендрограму іноді називають деревоподібною схемою, деревом об'єднання кластерів [3, с. 16]. У дендрограмі на вертикальній осі відображаються відстані, на горизонтальній – об'єкти класифікації.

На основі даних табл. 2 за допомогою STATISTICA 6.1 побудовано дендрограму класифікації регіонів Волинської області за основними показниками відтворення населення у 2014 р. Для цього в результатах ієрархічної класифікації натиснемо «Вертикальная дендрограмма» й отримаємо рис. 1.

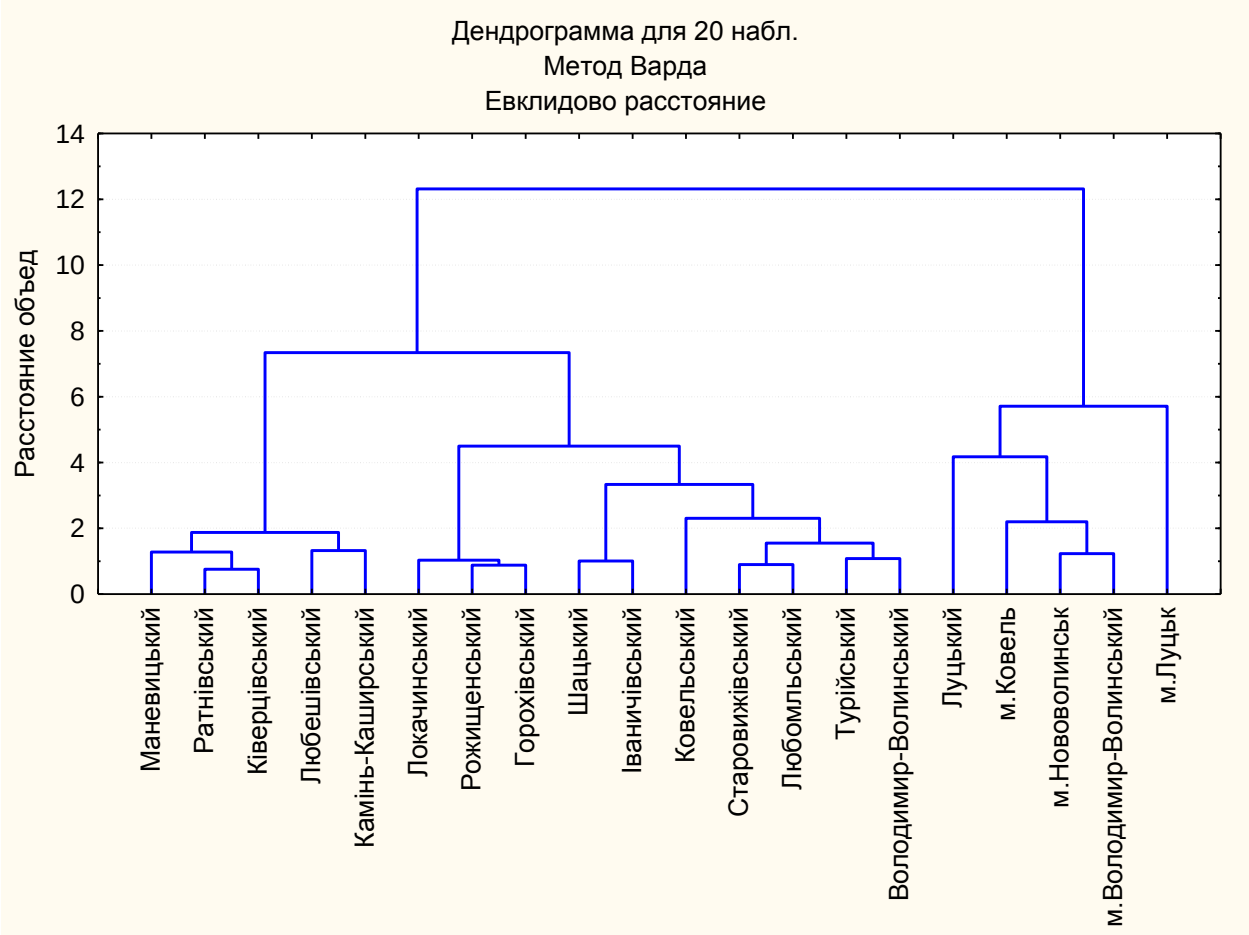

Рис. 1. *Дендрограма регіонів Волинської області за показниками відтворення населення у 2014 р.*

Вертикальна дендрограма читається зліва направо. Так, на рис. 1 наочно представлено три кластери. У перший кластер увійшли Маневицький (15), Ратнівський (16), Ківерцівський (9), Любешів-

ський (13), Камінь-Каширський (8), тобто райони з найвищими значеннями природного приросту населення. Так, у Камінь-Каширському районі рівень народжуваності перевищує середній по області на 3,8 осіб у розрахунку на 1000, або на 30 %, рівень смертності менший на 0,5 осіб у розрахунку на 1000, або на 3,8 %, що привело до максимально природного приросту в області (5,2 на 1000 населення). Райони цього кластера розміщені на півночі та сході Волинської області, на кордоні з Рівненською областю.

Другий кластер об'єднав 10 районів із найнижчими показниками відтворення населення: Володимир-Волинський (5), Горохівський (6), Іваничівський (7), Ковельський (10), Локачинський (11), Любомльський (14), Рожищенський (17), Старовижівський (18), Турійський (19), Шацький (20). Серед них можна виділити три райони з найбільшим природним убутком: Володимир-Волинський, природне скорочення в якому становить 4,1 особи в розрахунку на 1000; Горохівський – 3,9; Локачинський – 3,7 особи в розрахунку на 1000. У всіх районах другого кластера переважає сільське населення. Так, найменша частка сільських мешканців у Рожищенському районі становить 62 %, а найбільша – у Володимир-Волинському (91 %). Це призводить до того, що середній вік, питома вага осіб у віці 60 років і старше перевищує середнє значення по області.

У третій кластер (із середніми значеннями показників відтворення населення) увійшли всі міста обласного значення: Луцьк (1), Володимир-Волинський (2), Ковель (3), Нововолинськ (4) і Луцький район (12). У цьому кластері лише в Луцькому районі загальний коефіцієнт народжуваності перевищує середній по області на 7,1 % і становить 15,1 особи на 1000 населення, рівень смертності є нижчим за середній, лише в м. Нововолинську спостерігають природний убуток населення – 1,6 осіб у розрахунку на 1000.

Для більш детальної характеристики кожного кластера пакет STATISTICA 6.1 дає змогу розрахувати основні описові статистики.

У меню «Анализ-Основные статистики и таблицы» вибираємо «Группировка и Однофакторный ДА» [6], розрахунки проводимо на основі даних табл. 1. Отримуємо таблиці основних характеристик утворених кластерів для кожного показника: загального коефіцієнта народжуваності, загального коефіцієнта смертності, коефіцієнта фемінізації, кількості шлюбів, одиниць, загального коефіцієнта міграційного приросту (рис. 2).

Итоговая таблица средних (Таблица)
N=20 (Нет пропусков в завис. перем.)

Кластери	Загальний коефіцієнт народжуваності, ‰ Среднее	Загальний коефіцієнт народжуваності, ‰ N	Загальний коефіцієнт народжуваності, ‰ Ст.откл.
1	16,34000	5	1,105893
2	13,97000	10	1,428325
3	12,56000	5	1,974335
Всего	14,21000	20	1,996286

Итоговая таблица средних (Таблица)
N=20 (Нет пропусков в завис. перем.)

Кластери	Загальний коефіцієнт смертності, ‰ Среднее	Загальний коефіцієнт смертності, ‰ N	Загальний коефіцієнт смертності, ‰ Ст.откл.
1	13,46000	5	0,912688
2	16,29000	10	1,225153
3	10,98000	5	1,169188
Всего	14,25500	20	2,518453

Итоговая таблица средних (Таблица) N=20 (Нет пропусков в завис. перем.)			
Кластери	Коефіцієнт фемінізації, ‰ Среднее	Коефіцієнт фемінізації, ‰ N	Коефіцієнт фемінізації, ‰ Ст.откл.
1	1056,200	5	23,25296
2	1098,900	10	25,47526
3	1154,600	5	45,16968
Всего	1102,150	20	46,21605

Итоговая таблица средних (Таблица) N=20 (Нет пропусков в завис. перем.)			
Кластери	Кількість шлюбів, одиниць Среднее	Кількість шлюбів, одиниць N	Кількість шлюбів, одиниць Ст.откл.
1	341,200(	5	72,8231
2	179,500(	10	67,5265
3	714,200(	5	664,5805
Всего	353,600(	20	382,716(

Итоговая таблица средних (Таблица) N=20 (Нет пропусков в завис. перем.)			
Кластери	Загальний коефіцієнт міграційного приросту, ‰ Среднее	Загальний коефіцієнт міграційного приросту, ‰ N	Загальний коефіцієнт міграційного приросту, ‰ Ст.откл.
1	-1,42000	5	1,125611
2	-0,06000	10	2,320536
3	2,84000	5	2,842182
Всего	0,32500	20	2,657437

Рис. 2. *Основні характеристики кластерів*

Найбільший рівень смертності спостерігаємо в десяти районах другого кластера – 16,29 осіб на 1000 населення. У районах першого кластера середнє значення загального коефіцієнта смертності становить 13,46, у третьому – 10,98 осіб на 1000 населення. Перевищення рівня народжуваності в цих кластерах призвело до приросту населення в цілому по області. Найбільш однорідними за рівнем смертності є перша та третя групи.

Найкращу ситуацію щодо співвідношення жінок і чоловіків спостерігаємо в районах першого кластера, де на 1000 чоловіків доводиться 1056 жінок, другого кластера – 1099. Оскільки в третій кластер входять найбільші міста області, співвідношення жінок і чоловіків значно перевищує середнє по області. Найбільш однорідними за рівнем фемінізації є перший та другий кластери.

Найменше шлюбів у 2014 р. взяли мешканці районів другого кластера, у середньому 180 шлюбів на район. Цей кластер за цією ознакою найбільш однорідним. У зв'язку з тим, що в м. Луцьку зареєстровано найбільшу кількість шлюбів (1892), у третьому кластері спостерігаємо найбільшу середню кількість шлюбів (714) і найбільше стандартне відхилення.

У районах першої та другої груп простежено перевищення вибулих над прибулими, що призвело до від'ємного значення міграційного приросту. У третьому кластері відбується зростання чисельності населення на 2,84 особи в розрахунку на 1000 за рахунок міграції.

Отже, райони першого кластера утворюють найбільш однорідну групу з найменшими значеннями стандартних відхилень за чотирма з п'яти групувальних показників, виняток становить лише показник кількості шлюбів. Найбільш неоднорідною є група міст третього кластера, для якого стандартні відхилення майже всіх показників групування, крім загального показника смертності, є максимальними.

Висновки й перспективи подальших досліджень. Проведений аналіз регіональної диференціації демографічної ситуації у Волинській області на основі спеціалізованого програмного пакета STATISTICA 6.1 дав підставу виявити три основні кластери регіонів, які суттєво відрізняються за рівнем демографічних процесів. Так, найкраща демографічна ситуація у 2014 р. у п'яти районах півночі та сходу області, які увійшли до першого кластера, характеризується найвищими показниками природного приросту населення, найбільш оптимальним співвідношенням жінок і чоловіків, що сприяє підвищенню рівня шлюбності й народжуваності. Для третього кластера, у який увійшли чотири міста обласного значення та Луцький район, також характерний природний приріст населення, хоча значно менший, ніж у першому кластері. Найбільш складна демографічна ситуація в районах, що увійшли в другий кластер. За винятком Ковельського району, у 2014 р. у всіх цих районах спостерігали природне зниження населення, а несприятлива демографічна ситуація поєднується з невисоким рівнем економічного розвитку. Усе це засвідчує потребу приділити особливу увагу підвищенню рівня життя населення, покращенню його якості, соціально-економічному розвитку цих районів.

Використання програмного пакета STATISTICA для проведення кластерного аналізу забезпечує реалізацію різноманітних методів кластеризації, наочне графічне представлення результатів, які уможливлюють виявлення проблемних моментів та способів їх розв'язання в демографічних дослідженнях.

Джерела та література

1. Айвазян С. А. Прикладная статистика. Основы эконометрики. – Том 1 : Теория вероятностей и прикладная статистика : учеб. для вузов : в 2 т. / С. А. Айвазян, В. Мхитарян. –[2-е изд., испр.]. – М. : ЮНИТИ-ДАНА, 2001. – 656 с.
2. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: для профессионалов / В. Боровиков. – 2-е изд. (+CD). – СПб. : Питер, 2003. – 688 с. : ил.
3. Буреева Н. Н. Многомерный статистический анализ с использованием ППП STATISTICA / Н. Н. Буреева. – Нижний Новгород : [б. и.], 2007. – 112 с.
4. Демографічний щорічник: населення Волинської області 2014 / за ред. Л. С. Баранюк. – Луцьк : Голов. упр. статистики у Волин. області, 2015. – 110 с.
5. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл ; [пер. з англ. Е. З. Демиденко] ; под ред. А. Я. Боярского. – М. : «Статистика», 1977. – 128 с. : ил.
6. Електронний підручник по статистиці [Електроний ресурс] // StatSoft, Inc. – 2001. – Режим доступу : http://www.statsoft.ru/home/textbook/default.htm.

Бегун Светлана. Использование кластерного анализа для исследования демографической ситуации в регионе. В статье осуществлена кластеризация регионов Волынской области в 2014 г. на основе общих коэффициентов рождаемости и смертности, коэффициента феминизации, количества браков и общего коэффициента миграционного прироста. Показаны аналитические и графические возможности применения программного пакета STATISTICA 6.1 для проведения многомерной группировки при анализе демографических процессов отдельного региона. Построено дендрограму регионов Волынской области по показателям воспроизводства населения в 2014 г. Рассмотрены основные описательные статистики каждого сформированного кластера.

Ключевые слова: кластерный анализ, демографическая ситуация, воспроизводство населения, кластеризация, статистические методы, демографические процессы, кластеры, таксономия, распознавание образов.

Begun Svitlana. Using Cluster Analysis for the Research of the Demographic Situation in Regione. In the article the clusterization of regions of the Volyn region in 2014 was carried out on the basis of general coefficients of birth-rate and death rate, of coefficient of феминизации, of amount of marriages and of general coefficient of migratory increase. Analytical and graphical possibilities of application of programmatic package of STATISTICA 6.1 were shown for realization of multidimensional groupment at the analysis of demographic processes of separate region of country. Dendrogramof the regions of Volyn region in terms of reproduction of the population in 2014 was built. The main descriptive statistics for each formed cluster were described.

Key words: cluster analysis, demographics, population reproduction, clustering, statistical methods, demographic processes, cluster, taxonomy, pattern recognition.