В. С. Грицевич

Львівський національний університет імені Івана Франка, gvsmg@ukr.net

МЕТОДИЧНІ ОСНОВИ ВИВЧЕННЯ ДЕМОГЕОГРАФІЧНОЇ САМООРГАНІЗАЦІЇ РЕҐІОНУ В СВІТЛІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

Поняття про самоорганізацію і синергетику. Останнім часом зростає інтерес науковців різних галузей знання до такого явища, як самоорганізація. Виник міждисциплінарний науковий напрямок *синергетика*, який спеціалізується на дослідженні ефектів самоорганізації. Така ситуація є дуже цінною для суспільної географії. Її предметом дослідження [8,9] є геопросторова організація суспільства (або геопросторова організація геопросторово мінливого суспільного світу [4]). З такої точки зору в складі суспільної географії можна розглядати *суспільно-географічну синергетику* з предметом дослідження *геопросторова самоорганізація суспільства*. Більше того, оскільки переважна частина суспільних явищ розвиваються автономно під дією суспільних рушійних сил, то виявляється, що майже вся суспільна географія є географією самоорганізації суспільства.

Оскільки при самоорганізації суспільства створюються нові суспільні структури,то у філософському плані процес самоорганізації можна розглядати як перехід від небуття до буття. В результаті виникають суспільно-географічні утворення: таксони, кластери, райони, агломерації і т.д. У синергетиці важлива роль відводиться поняттю атрактора, тобто такого реального чи гіпотетичного стану досліджуваного об’єкта, який “притягує” до себе інші стани. Ми будемо визначати атрактор, як центр ваги відповідного таксону (чи іншого самоорганізаційного утворення). Тому ефект “притягання” розглядаємо як специфічну “функцію місця” для атрактора.

Поняття про інтелектуальний аналіз даних. Останні десятиліття стали роками бурхливого розвитку технологій інтелектуального аналізу даних. Йому присвячена велика кількість публікацій та Інтернет-сайтів [1,6,7]. Ці технології є універсальними і можуть бути використані в будь-якій галузі наукових досліджень. Тому, цілком природньою є спроба застосувати їх до потреб суспільної географії. Інтелектуальний аналіз даних спрямований на те, щоб на основі масиву спостережень отримати змістовні знання. Масив спостережень, як правило, має вигляд сукупності статистичних таблиць, у кожній з яких подані значення певних показників у розрізі територіальних елементів. Такий масив є, за своєю суттю, географічним, однак він містить голі числові дані і в ньому не видно безпосередньо змістовних зв’язків між показниками.

Інтелектуальний аналіз даних - це процес виявлення в сирих даних раніше невідомих, нетривіальних, практично корисних і доступних для інтерпретації знань, необхідних для ухвалення рішень в різних сферах людської діяльності, це мультидисциплінарна область, що виникла і розвивається на базі таких наук як прикладна

статистика, розпізнавання образів, штучний інтелект, теорія баз даних і ін. Можна сказати ще так: інтелектуальний аналіз даних, це аналіз, побудований з використанням засобів штучного інтелекту. Він включає підходи, об'єднані ідеєю використання теорії штучного інтелекту.

До цих підходів відносять такі методи: штучні нейронні мережі; еволюційне програмування (в т.ч. алгоритми методу групового врахування аргументів); генетичні алгоритми; асоціативна пам'ять (пошук аналогів, прототипів); нечітка логіка; дерева рішень; системи обробки експертних знань та інші.

Приклади дослідження демогеографічної ситуації в регіоні із застосуванням таких методів інтелектуального аналізу даних, як "дерево рішень" та "нейромережі Кохонена" приведені в [3,5].

Одним із цікавих і важливих методів інтелектуального аналізу даних є аналіз асоціацій. Нас, у контексті цього дослідження, в першу чергу буде цікавити пошук саме *територіальних асоціацій*, як самоорганізаційних утворень.

Нехай *TotalNumberofTransaction* - загальна зареєстрована кількість транзакцій, A,B - елементи транзакцій. Через $AbsoluteSupport(A)$, $AbsoluteSupport(B)$, $AbsoluteSupport(A,B)$ позначаємо відповідно міру абсолютної підтримки (кількість разів появи в загальній сукупності транзакцій) елементів A, B та їхнього поєднання A,B.

На основі абсолютних підтримок обчислюємо відносні підтримки.

Міри відносної підтримка елементів A та B:

$$RelativeSupport(A) = \frac{AbsoluteSupport(A)}{TotalNumberofTransaction}.$$

$$RelativeSupport(B) = \frac{AbsoluteSupport(B)}{TotalNumberofTransaction}.$$

Міра відносної підтримки поєднання елементів A,B:

$$RelativeSupport(A,B) = \frac{AbsoluteSupport(A,B)}{TotalNumberofTransaction}.$$

Для результуючих асоціативних правил обчислюють міри їхньої вірогідності (*Probability*) та важливості (*Importance*).

Для правила $A \to B$:

$$Probability(A \to B) = Probability(B/A) = \frac{AbsoluteSupport(A,B)}{AbsoluteSupport(A)};$$

$$Importance(A \to B) = \frac{Probability(A \to B}{Probability(\bar{A} \to B)}.$$

Для правила $B \to A$:

$$Probability(B \to A) = Probability(A/B) = \frac{AbsoluteSupport(A,B)}{AbsoluteSupport(B)};$$

$$Importance(B \to A) = \frac{Probability(B \to A)}{Probability(\bar{B} \to A)}.$$

Інколи додатково обчислюють логарифм від міри важливості. Це дає змогу робити висновки на основі знаку отриманого числа.

Математико-географічна характеристика територіальних самоорганізаційних структур. Розглянемо регіон, як територіальний комплекс, що складається з M територіальних об'єктів (субрегіонів, адміністративних районів і т.д.). Об'єкти характеризуються N ознаками, так що Z_{ij} - значення j-ї абсолютної ознаки на i-му об'єкті (тут і далі $i = 1,\ldots,M$; $j = 1,\ldots,N$). Позначимо через x_i, y_i координати центра ваги i-го об'єкта, а через S_i, P_i - відповідно його площу і значення деякої обсягової характеристики (наприклад чисельність населення), яку використовуємо для переведення абсолютних ознак у відносні за формулою: $z_{ij} = \frac{Z_{ij}}{P_i}$. Отримані відносні ознаки нормалізуємо за формулою: $w_{ij} = \frac{z_{ij} - z_i}{\sigma_i}$, де $z_i = \frac{1}{M}\sum_{i=1}^{M} z_{ij}$, $\sigma_i = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(z_{ij} - z_i)^2}$.

У результаті самоорганізації, об'єкти регіону об'єднуються в таксони. Позначимо через T_k множину номерів об'єктів k-го таксона. Вважаємо, що $\bigcup T_k = \{1,2,\ldots,M\}$. Нехай E_k - кількість об'єктів k-го таксона, тобто $E_k = \sum_{i \in T_k} 1$, $E_k = M$.

Спочатку, для кожного k-го таксона визначаємо координати (X_S^k, Y_S^k) його територіального центра ваги:

$$X_S^k = \frac{\sum_{i \in T_k} S_i \cdot x_i}{\sum_{i \in T_k} S_i}, \quad Y_S^k = \frac{\sum_{i \in T_k} S_i \cdot y_i}{\sum_{i \in T_k} S_i},$$

а також координати (X_P^k, Y_P^k) центра ваги обсягової ознаки (населення):

$$X_P^k = \frac{\sum_{i \in T_k} P_i \cdot x_i}{\sum_{i \in T_k} P_i}, \quad Y_P^k = \frac{\sum_{i \in T_k} P_i \cdot y_i}{\sum_{i \in T_k} P_i}.$$

Далі визначаємо координати (X_j^k, Y_j^k) центра ваги k-го таксона за кожною j-ю абсолютною ознакою:

$$X_j^k = \frac{\sum_{i \in T_k} Z_{ij} \cdot x_i}{\sum_{i \in T_k} Z_{ij}}, \quad Y_j^k = \frac{\sum_{i \in T_k} Z_{ij} \cdot y_i}{\sum_{i \in T_k} Z_{ij}}.$$

$i\ T_k^{ij}$ $i\ T_k^{ij}$

Сукупність N центрів ваги $(X_1^k, Y_1^k), \ldots, (X_N^k, Y_N^k)$ утворює своєрідну "плеяду" для кожного k-го таксона. Їхнє взаємне розміщення характеризує особливості стану об'єктів таксона. Метрична інформація про таке взаємне розміщення міститься в матриці географічних відстаней $\{D_{jl}^k\}$ для точок "плеяди", де $D_{jl}^k = \sqrt{(X_j^k - X_l^k)^2 + (Y_j^k - Y_l^k)^2}$. Важливою характеристикою взаємного розміщення цих центрів ваги є також найцентральніша ознака таксона, яку визначаємо з умови близькості до територіального центра ваги: $\min_{1 \le j \le N} \sqrt{(X_j^k - X_S^k)^2 + (Y_j^k - Y_S^k)^2}$.

В N-вимірному просторі нормалізованих відносних ознак кожен таксон також має свій центр ваги з координатами $(W_1^k, W_2^k, \ldots, W_N^k)$, які обчислюються за формулою $W_j^k = \frac{1}{E_k} \oplus_{i\ T_k} w_{ij}$. Ці центри ваги характеризують особливості кожного k-го стану і є для цього стану вузлами притягання.

Структура демогеографічних даних регіону. Для ефективного аналізу демогеографічної самоорганізації потрібна добре вибрана і структурована система даних, яка дає змогу з різних боків охарактеризувати демогеографічну ситуацію регіону. Така система головних демогеографічних ознак представлена на рис. 1. Система складається з двох підсистем: ознак структури населення і ознак руху населення.

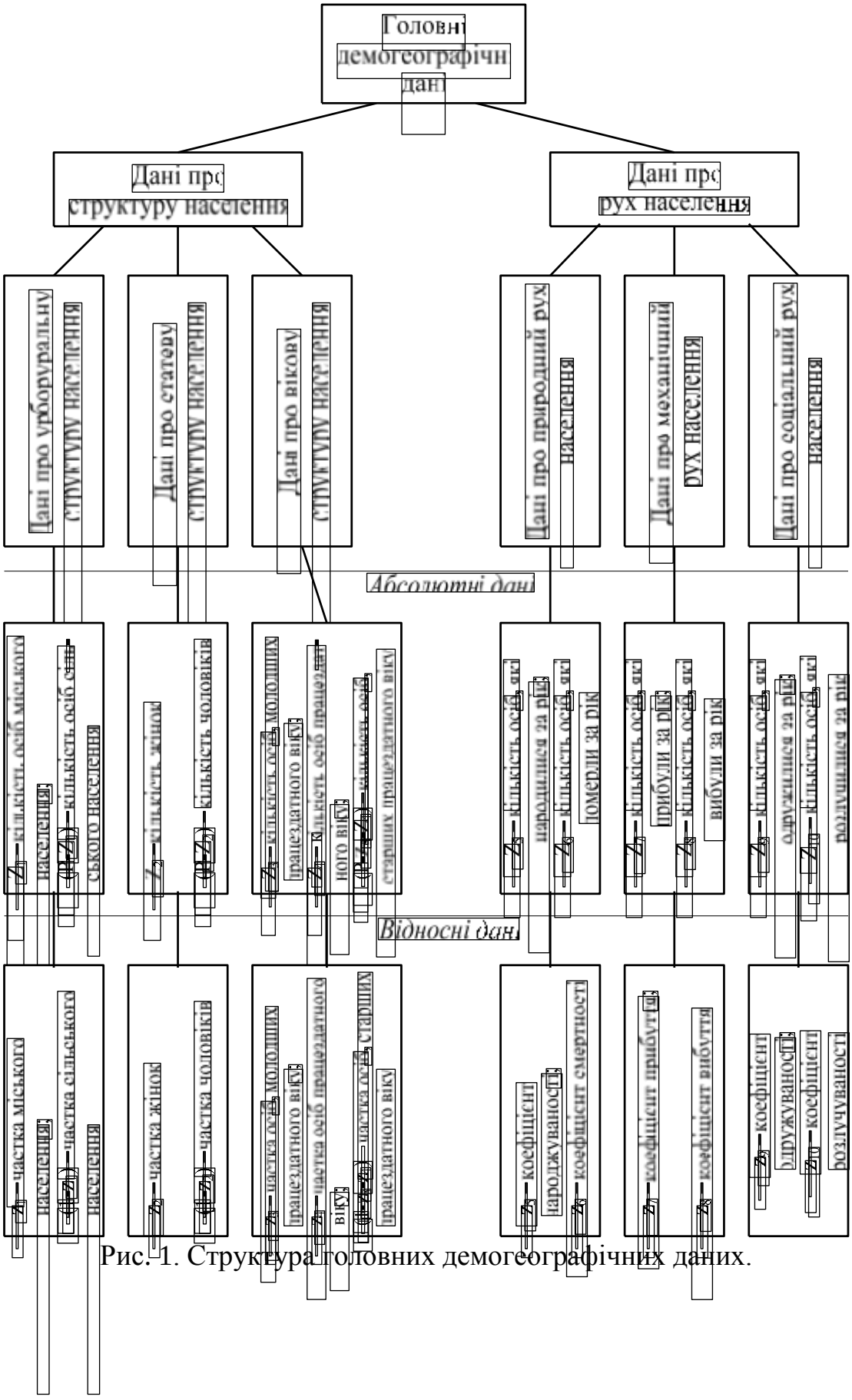

Рис. 1. Структура головних демогеографічних даних.

Абсолютні дані про урбораль ну структуру населення включають кількість осіб міського та сільського населення. Оскільки ці кількості пов'язані між собою (їхня сума рівна загальній кількості населення), то ми описуємо таку структуру одною незалежною ознакою Z_1. У рамках центрографічного аналізу становить інтерес дослідження центрів ваги як міського, так і сільського населення. При цьому слід пам'ятати, що, в силу залежності відповідних ознак, центр ваги усього населення $\left(X_P^k, Y_P^k\right)$ лежатиме на відрізку прямої лінії, яка з'єднує центри ваги міського та сільського населення.

При переході до відносних ознак отримуємо частку міського $z_1 = \frac{Z_1}{P}$ і частку сільського $\frac{P - Z_1}{P}$ населення. Ці частки також пов'язані між собою і їхня сума рівна одиниці[1].

Абсолютні дані про статеву структуру населення включають кількість жінок і кількість чоловіків. Оскільки ці кількості пов'язані так само, як у попередньому випадку, то ми описуємо цю структуру одною незалежною змінною Z_2. У рамках центрографічного аналізу становить інтерес дослідження центрів ваги як жінок так і чоловіків. При цьому слід пам'ятати, що, в силу залежності відповідних ознак, центр ваги усього населення $\left(X_P^k, Y_P^k\right)$ лежатиме на відрізку прямої лінії, яка з'єднує центри ваги жінок та чоловіків. При переході до відносних ознак отримуємо частку жінок $z_2 = \frac{Z_2}{P}$ і частку чоловіків $\frac{P - Z_2}{P}$. Сума цих часток також рівна 1.

Абсолютні дані про вікову структуру населення включають дані про кількість осіб молодших працездатного віку, кількість осіб працездатного віку, кількість осіб старших працездатного віку. Ці три кількості пов'язані між собою (їхня сума рівна загальній кількості населення). Тому ми описуємо таку структуру двома незалежними змінними Z_3 і Z_4. У рамках центрографічного аналізу дуже цікавим є дослідження центрів ваги усіх трьох головних вікових груп. При цьому слід пам'ятати, що, в силу залежності відповідних ознак, центр ваги усього населення $\left(X_P^k, Y_P^k\right)$ лежатиме всередині трикутника, утвореного центрами ваги трьох головних вікових груп. При переході до відносних ознак отримуємо відповідно три частки $z_3 = \frac{Z_3}{P}$, $z_4 = \frac{Z_4}{P}$, $\frac{P - Z_3 - Z_4}{P}$ сума яких рівна 1.

Абсолютні дані про природний рух населення включають дані про кількість осіб, які народилися Z_5 та кількість осіб, які померли Z_6[1]

Тут і далі розглядаємо відносні ознаки структури населення, обчислені для потреб подальшої математичної обробки. Коли їх обчислюють для ручного аналізу, ці частки домножують на 100 і вимірюють у відсотках.

на протязі року. Ці змінні є незалежними, тому, при потребі, одну з них (наприклад Z_6) можна замінити на різницю $(Z_5 \quad Z_6)$, що виражає величину природного приросту населення. При переході від Z_5, Z_6 до відносних ознак отримуємо $z_5 = \frac{Z_5}{P}$ - коефіцієнт народжуваності населення, $z_6 = \frac{Z_6}{P}$ - коефіцієнт смертності населення[1], $(z_5 \quad z_6)$ - коефіцієнт природного руху населення.

Абсолютні дані про механічний рух населення включають дані про кількість осіб, які прибули Z_7 та кількість осіб, які вибули Z_8 на протязі року. Ці змінні є незалежними, тому, при потребі, одну з них (наприклад Z_8) можна замінити на різницю $(Z_7 \quad Z_8)$, що виражає величину міграційного приросту населення. При переході від Z_7, Z_8 до відносних ознак отримуємо $z_7 = \frac{Z_7}{P}$ - коефіцієнт прибуття населення і $z_8 = \frac{Z_8}{P}$ - коефіцієнт вибуття населення, $(z_7 \quad z_8)$ - коефіцієнт механічного (міграційного) руху населення.

Абсолютні дані про соціальний рух населення (на прикладі сімейно-шлюбних процесів) включають дані про кількість осіб, які одружилися Z_9 та кількість осіб, які розлучилися Z_{10} на протязі року. При переході від Z_9, Z_{10} до відносних ознак отримуємо $z_9 = \frac{Z_9}{P}$ - коефіцієнт одружуваності населення і $z_{10} = \frac{Z_{10}}{P}$ - коефіцієнт розлучуваності населення.

Таксономічний аналіз демогеографічної самоорганізації регіону. При вивченні демогеографічної самоорганізації регіону методами таксономізації починаємо з матриці спостережень $\{Z_{ij}\}$ розміром $M \cdot N$ як і раніше. На її основі обчислюємо матрицю відносних ознак $\{z_{ij}\}$, а далі матрицю нормалізованих ознак $\{w_{ij}\}$. Для останньої матриці обчислюємо матрицю таксономічних відстаней в N-вимірному просторі нормалізованих відносних ознак і застосовуємо відомі процедури багатовимірної таксономізації (наприклад метод "дерева поєднань") з відповідною картографічною реалізацією [2]. Отримані таксони розглядаємо, як самоорганізаційні утворення, кожен з яких представляє певний можливий стан територіального об'єкта.

У випадку демогеографічної самоорганізації розгядаємо вищеописані ознаки, що характеризують населення. Для першого блоку ознак $N = 4$, а для другого блоку $N = 6$.

Якщо розглядати Львівську область у розрізі адміністративних

[1] Тут і далі розглядаємо відносні ознаки руху населення, обчислені для потреб подальшої математичної обробки. Коли їх обчислюють для ручного аналізу, ці частки домножують на 1000 і вимірюють у проміле.

районів (міста обласного підпорядкування включаємо у відповідні райони), а також окремо м. Львів, то $M = 21$.

Асоціативний аналіз демогеографічної самоорганізації регіону. У цьому випадку початково також маємо матрицю спостережень $\{z_{ij}\}$. Однак, тепер спочатку переводимо значення ознак з неперервної шкали в дискретну. Для цього можна скористатись дискретними шкалами з різною кількістю рівнів. Найпростішою є шкала, що має три рівні значення ознаки: "Низький" (*Н*), "Середній" (*С*), "Високий" (*В*). Якщо потрібно точніше зафіксувати градації ознаки, то можна скористатись п'ятирівневою шкалою з такими рівнями: "Низький" (*Н*), "Нижче середнього" (*НС*), "Середній" (*С*), "Вище середнього" (*ВС*), "Високий" (*В*). При значних варіаціях значень на границях інтервалу мінливості ознаки існує ще така п'ятирівнева шкала: "Дуже низький" (*ДН*), "Низький" (*Н*), "Середній" (*С*), "Високий" (*В*), "Дуже високий" (*ДВ*).

Розглянемо трирівневу шкалу з рівнями $\{H, C, B\}$. У цій задачі транзакцією буде спостереження за рівнями ознак одного територіального об'єкта регіону, тобто одна транзакція має вигляд:

$1РівеньОзнаки, 2РівеньОзнаки, \ldots, NРівеньОзнаки$

$$\text{де } jРівеньОзнаки = \begin{cases} jH, & \text{якщо рівень } j \text{ ї ознаки низький} \\ jC, & \text{якщо рівень } j \text{ ї ознаки середній} \\ jB, & \text{якщо рівень } j \text{ ї ознаки високий} \end{cases}.$$

Всього таких транзакцій буде стільки, скільки є спостережень за територіальними об'єктами, тобто M.

У результаті застосування методу асоціацій до цієї сукупності транзакцій отримаємо таксони територіальних об'єктів, у кожному з яких спостерігається певна асоціація (поєднання) рівнів деяких ознак, тобто побачимо самоорганізацію регіону за асоціацією рівнів демогеографічних ознак.

Висновки. Головною частиною предмету дослідження суспільної географії є геопросторова самоорганізація суспільства в різних її формах. Сучасна суспільна географія разом із синергетикою впритул підійшли до вивчення геопросторових суспільних самоорганізаційних утворень. Для таких досліджень, поряд із традиційними методами, варто застосувати нові, зокрема методи інтелектуального аналізу даних. При вивченні таксонів, як самоорганізаційних утворень, за їхні атрактори приймаємо (у першому наближенні) відповідні центри ваги. Загальну демогеографічну характеристику регіону доцільно здійснювати за двома головними блоками даних: ознаками структури населення і ознаками руху населення. Для виявлення самоорганізаційних утворень використовуємо як традиційний багатовимірний аналіз так і метод асоціацій з арсеналу інтелектуального аналізу даних з подальшою картографічною інтерпретацією отриманих результатів.

Список використаних джерел:

1. Анализ данных и процессов / Барсегян А.А. и др. СПб: БХВ-Петербург, 2009.
2. Грицевич В.С. Картографічна інтерпретація багатовимірної таксономізації в суспільно-географічних дослідженнях // Вісник геодезії та картографії. №3, 2005. –С.34-38.
3. Грицевич В.С. Суспільно-географічні застосування інтелектуального аналізу демографічних даних на прикладі Львівського регіону // Матеріали III Всеукр. наук.-практ. конф. «Географія та екологія: наука та освіта». – Умань, 2010. –С.49-51.
4. Грицевич В.С. Головні категорії та поняття суспільно-географічного пізнання дійсності // Часопис соціально-економічної географії. -Вип. 9(2). - Харків, 2010. -С.19-24.
5. Грицевич В.С. Застосування нейромереж Кохонена для демографічної таксономізації регіону // Проблеми розвитку прикордонних територій та їх участі в інтеграційних процесах: Матеріали VII Міжнар. наук.-практ конф. –Луцьк: Волинський нац. ун-т ім. Лесі Українки. 2010. –С.435-438.
6. Дюк В., Самойленко А. Data Mining: учебный курс. - СПб: Питер, 2001. - 368 с.
7. Макленнен Дж., Танг Чж., Криват Б. Microsoft SQL Server 2008: Data Mining – интеллектуальный анализ данных: Пер. з англ. – Санкт-Петербург: БХВ-Петербург, 2009.
8. Топчієв О.Г. Основи суспільної географії. –Одеса: Астропринт, 2009.
9. Шаблій О.І. Основи загальної суспільної географії. Підручник. –Львів: Видавничий центр ЛНУ ім. Івана Франка, 2003.