УДК 681.325

В.В. Путренко

Геостатистичний інтелектуальний аналіз геохімічних даних

Національний технічний університет України
«Київський політехнічний інститут»

Розглянуто методику геостатистичного інтелектуального аналізу геохімічних даних, яка дозволяє в автоматизованому режимі здійснювати аналіз геопросторових закономірностей поширення хімічних елементів і побудови електронних і паперових карт. Визначено основні етапи проведення геостатистичного аналізу даних і розкрито підходи до континуального подання даних засобами векторних і растрових моделей. Виконано порівняльний аналіз переваг і недоліків інтерполяції геохімічних даних різними методами.
Ключові слова: інтелектуальний аналіз даних, ГІС, геостатистика, інтерполяція, геохімія

Вступ. Фундаментальними засадами розвитку навколишнього середовища є його континуальність і безперервність, що дозволяє моделювати просторовий розподіл явищ на основі припущення їх поступової зміни. Метод побудови геопросторових полів, який ґрунтовно методологічно був розроблений в роботах зарубіжних і вітчизняних учених, зараз широко застосовується для моделювання географічних закономірностей розподілу показників та явищ. Особливо важливе значення метод має для моделювання рельєфу та його морфометричних показників. Завдяки застосуванню ГІС-технологій підготовка моделей, їх побудова та картографування стали частиною єдиного автоматизованого процесу оброблення та зберігання геопросторової інформації.

Геостатистичне моделювання полів розподілу геохімічних показників є прикладом переваги просторового підходу над іншими. Без аналізу розподілу просторових трендів у поширенні хімічних елементів неможливо проведення заходів з раціонального природокористування. Розподіл геохімічних показників дозволяє виявляти поклади корисних копалин, знаходити аномальні значення вмісту елементів, які можуть впливати на життєдіяльність людини, проводити ландшафтні дослідження.

Вихідні передумови. У галузі інтелектуального аналізу даних в Україні широко відомі роботи Інституту прикладного системного аналізу НТУУ «КПІ», Світового центру даних з геоінформатики та сталого розвитку М.З. Згуровського, Н.Д. Панкратової, А.О. Болдака [1 – 2].

Питання геостатистичних досліджень розглянуто в зарубіжних роботах Є.П. Ісаакса, Р.М. Шрівастава, Н.В. Кларка, Х. Вакернагеля, Р. Уебстера, М.О. Олівера [3 – 6].

Дослідженням картографічного моделювання геохімічних полів присвятили свої роботи В.Т. Жуков, Б.А. Новаковський, А.Н. Чумаченько, В.Г. Лінник, О.Р. Мусин, С.Н. Сербенюк та ін. [7 – 10].

Формулювання мети статті, постановка задачі. Метою дослідження є опрацювання методики геостатистичного інтелектуального аналізу геохімічних даних. Задачами дослідження є визначення чинників, що впливають на вибір методів геостатистичного аналізу, основних етапів моделювання геохімічних даних і способів укладання остаточних електронних картографічних творів.

Виклад основного матеріалу. Апробацію методики було здійснено на основі даних про хімічний склад підземних вод на території України за вмістом урану, миш'яку та фторидів, як речовин, що можуть мати негативний вплив на споживчу якість води у разі перевищення їх концентрації [10].

Першим етапом підготовки до проведення аналізу геохімічної інформації є збір первинних даних. На цьому етапі визначають схему відбору проб, яка залежить від мети дослідження, особливостей розподілу досліджуваних елементів, статистичної моделі, яку приймають за основу, фінансування (рис. 1).

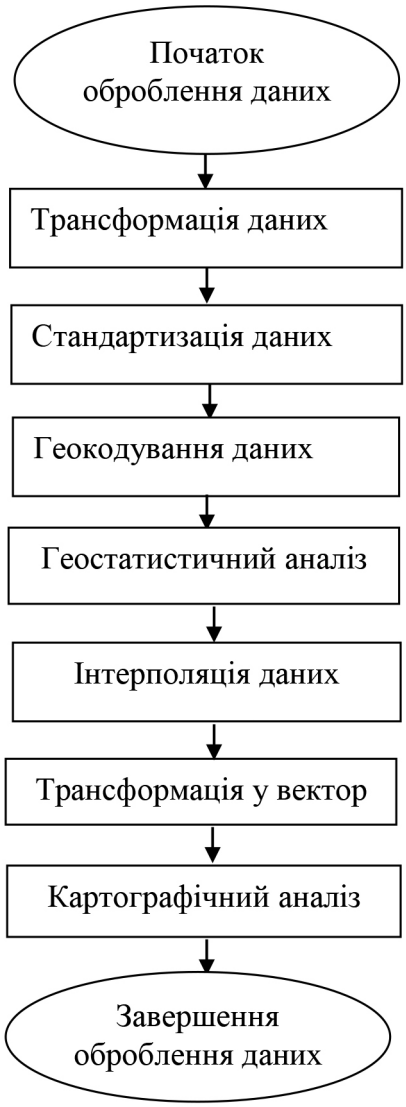

Рис. 1. Етапи геостатистичного моделювання геохімічних даних

Дослідження проводили на всій території країни. За точки відбору приймали місця впадіння рік, тому структура мережі точок є нерегулярною. На кожній точці було визначено вміст елементів у воді та встановлено географічні координати. Результатом роботи стала реляційна база даних з точками відбору води та показниками вмісту елементів.

Другим етапом є трансформація та стандартизація табличної бази даних з метою перетворення на геопросторову. Для цього використовують процедуру

геокодування, яка вбудована у більшість сучасних ГІС-продуктів. У результаті геокодування створюється база геоданих, що містить точкові об'єкти у вибраній системі координат. Ці дані є первинними для створення моделі геохімічних даних.

Наступний етап передбачає побудову геостатистичної моделі розподілу даних для проведення аналізу. Існує дві основні моделі полів: растрова та векторна. Растрову модель поділяють за видами інтерполяції даних та їх аналізу, векторну – за видами векторних об'єктів: точкові, лінійні, полігональні та TIN-модель. Залежно від вихідних даних, мети оброблення даних, точності та можливостей аналізу даних вибирають модель подання інформації. У більшості випадків спочатку створюють растрові моделі, а на їх основі – векторні.

Останній етап містить роботи з картографічного аналізу. До них відноситься вибір базових геопросторових даних, які відповідають деталізації певного масштабу, форми подання картографічних даних у вигляді растрового чи векторного відображення, розміщення у форматі ГІС, електронного атласу чи публікація в Інтернеті, макетування. Результатом є отримання карти або серії карт, які описують розподіл геохімічних показників.

За основу для створення геостатистичних моделей геохімічних даних були вибрані дані державного підприємства «Кіровгеологія», які містять відомості про вміст урану, миш'яку, фторидів у ґрунтових водах. База даних була створена на основі збору результатів геологічної зйомки у місцях впадіння рік. Проби відбирали у літній період, коли основним джерелом живлення річок є ґрунтові води. Масштаб зйомки дорівнює 1: 1 000 000. Точність координат прив'язки точок відбору проб відповідає масштабу 1: 200 000. База містить результати аналізу проб води у 6550 точках в Україні, а також за її межами на території Росії, Білорусі, Молдови. Введемо такі позначення: $\{A\} = \{A_1, \ldots, A_n\}$ – множина точок відбору проб.

Вона складається з двох таблиць. Перша таблиця містить дані про умови відбору проб: відповідальна група, дата, країна, геологічна область, координати відбору проби у прямокутній та географічній системі координат. Друга таблиця містить відомості про кількісний вміст хімічних елементів.

У багатьох випадках під час проведення зйомок використовують регулярну мережу досліджень, основану на сітці квадратів, що рівномірно покривають усю територію.

У даному випадку при проведенні зйомки використовували нерегулярну мережу, пов'язану з елементами річкової мережі. Вона має різну щільність розміщення точок, що пов'язано зі зміною щільності річкової мережі та збільшенням кількості точок відбору у місцях прогнозування родовищ корисних копалин. Картографічне моделювання засобами ГІС дозволяє уникнути проблем, які виникають при проведенні дослідження на основі нерегулярної мережі та в умовах динамічної зміни щільності мережі.

Геокодування є головною процедурою переходу від реляційної структури даних до геореляційної. У результаті цього ми отримуємо об'єкти, розміщені в просторі $A\ (x,y)$ та придатні для геопросторового аналізу. Основними етапами процедури геокодування є визначення способу геокодування, проекції перетворення геопросторових даних, отримання географічних даних. Виділяють три основні види геокодування: на основі наявних координат, на основі суміщення атрибутивних даних та на основі суміщення геометричних даних. У першому випадку виявляються стовпці таблиці, які містять координатну інформацію. Ці дані можуть вирізнятись за способом отримання (супутникові системи позиціонування, інструментальні зйомки, вимірювання за картами), кількістю координатних вимірів

(широта, довгота, висота), системою координат (прямокутна, полярна, географічна), картографічною проекцією.

Суміщення атрибутивних даних дозволяє геокодувати дані на основі однакових значень атрибутів у таблиці з даними із таблиці з геометричними об'єктами. Цей спосіб геокодування у більшості випадків є менш точним.

Суміщення за геометричними даними дозволяє отримати достатньо точний результат. В основі процедури лежать принципи геометричної прив'язки та афінних перетворень даних, коли один геометричний шар прив'язується до іншого на основі ключових точок.

У випадку оброблення геологічної бази використано координати широти і довготи у географічній системі координат проекції Пулково 1942 для топографічних карт. Результати геокодування перевіряють під час суміщення з шарами географічної основи. Збіжність вузлів гідрографічної мережі з точками забору проб підтверджує правильність проведення процедури геокодування.

Під час геокодування було виявлено точки, які лежать далеко за межами загальної мережі спостережень. Координати цих точок вважаються помилковими та вилучаються із загальної вибірки. $A \subset B\{x_i\, y_i\, z_i\}$, де B задано екстентом області дослідження i -того набору точок.

Наступним кроком в опрацюванні даних є їх інтерпретація. Для цього будується безперервний розподіл значень у просторі на основі методів інтерполювання. Сучасні геоінформаційні системи ArcGis мають у своєму складі широкий арсенал засобів геостатистичного моделювання даних на основі методів інтерполяції [11]. До цих методів відноситься побудова TIN-моделей, метод середнього зважування обернено пропорційно відстані, сплайн, кригінг.

Статистичні поверхні можна моделювати на основі інтерполяції точкових полів або побудови TIN-моделей. Кожна модель має свої переваги та недоліки. Растр – більш проста модель поверхні, TIN-моделі можуть надавати більш точне уявлення поверхонь і просторових об'єктів, але потребують більших зусиль зі збору інформації.

При аналізі геохімічних даних із різною щільністю дослідної мережі доцільнішим є використання растрових поверхонь, які на відміну від TIN-моделей можуть надавати більш загальну картину розподілу вмісту елементів. Основним критерієм при цьому виступає відтворення загальних рис геохімічних полів на відміну від моделювання рельєфу. Важливе значення має роздільна здатність растрової моделі даних. Вона визначає точність проведення інтерполяції, подальший масштаб вихідних даних. Вибір розміру комірок растра залежить від площі території, кількості й щільності точок знімання, масштабу базових наборів геоданих [12].

TIN – нерегулярна тріангуляційна мережа, що складається з точок, кожна із яких зіставляється зі значенням. Моделювання геохімічних поверхонь потребує використання як значень точок характеристик вмісту хімічних елементів. За цими даними виконується побудова мережі трикутників (тріангуляція), яка створює безперервну поверхню у тривимірному просторі. Тріангуляція створює набір трикутників без перекриття (граней), які повністю заповнюють задану область.

Метод середнього зважування обернено пропорційно відстані є видом растрової інтерполяції даних. На його основі обчислюють значення комірок за середнім від суми значенням точок замірів, що знаходяться поблизу кожної комірки. Чим ближче точка до центра оцінюваної комірки, тим більшу вагу має її значення у процесі обчислення середнього від суми значенням точок замірів. Цей

метод передбачає, що вплив значень виміряної змінної зменшується пропорційно збільшенню відстані від точки замірювання

$$z(x_j) = \sum_{i=1}^{n} z(x_i) d_{ij}^{-r} / \sum_{i=1}^{n} d_{ij}^{-r} \quad (1)$$

де x_j – точки (вузли), для яких має бути інтерпольована поверхня, а x_i - точки з відомими значеннями; d_{ij}, – відстані («дистанції») між точками з відомими значеннями і точкою оцінювання; r – показник ступеня; n – кількість точок із відомими значеннями, що потрапляють в окіл вузла оцінювання.

Отримані значення комірок можуть контролюватися зміною ступеня впливу на комірку сусідніх точок, зміною радіуса пошуку та встановленням бар'єрів. Інтерполяція цим способом геохімічних полів дозволяє отримати дані про загальний розподіл геохімічних полів, але точність значень комірок растра буде відрізнятися від вхідних значень.

На основі методу сплайнів розраховують значення комірок на основі математичної функції, що мінімізує кривизну поверхні, вираховують найбільш рівну поверхню, яка точно проходить через усі точки вимірів. Сплайни розраховують на основі методів регуляризації, коли створюється більш плавна поверхня, що може виходити за межі діапазону замірів, та методу натяжіння, коли поверхня найбільше наближена до існуючих значень:

$$z(x_0) = \sum_{i=1}^{n} c_i B(h_{0i}), \quad (2)$$

де h_{0i} – відстань від точки x_0 до точки x_i; $B(h_{0i})$ – базисна функція, що визначається від відстані; c_i – вагові коефіцієнти. Коефіцієнт c_i визначає алгебричний знак входження відповідного члена і ступінь його впливу. Класичний варіант методу є точним, але можливо введення параметра згладжування δ.

Зазвичай використовують такі типи базисних ядерних функцій:

природний кубічний сплайн: $B(h)=(h^2+\delta^2)^{1/2}$; (3)

тонкий сплайн: $B(h) = (h^2 + \delta^2)\lg(h^2 + \delta^2)$. (4)

Для налаштування інтерполяції методом сплайну застосовують параметри ваги та кількості точок, що беруть участь у розрахунках. Побудова сплайнів на основі регуляризації дозволяє отримати більш загальну картину розподілу хімічних елементів, на основі натяжіння – чіткі територіальні абриси з більш точними значеннями комірок.

Метод кригінгу відноситься до групи геостатистичних методів, основаних на геомоделях, що містять автореляцію (статистичний зв'язок між виміряними точками). Тому цей спосіб дозволяє не тільки отримати розрахункову поверхню, але визначити значення точності чи достовірності розрахунку.

Для розрахунку за методом кригінгу необхідно виявити правила залежності і розрахувати прогнозні значення. При вирішенні цих задач створюються варіограми та коваріаційні функції для оцінювання значення статистичних залежностей (просторової автокореляції) і визначаються прогнозні значення пустих комірок. Значення у точках складаються з таких компонентів:

$$z(x)=m(x)+\varepsilon'(x)+\varepsilon'', \quad (5)$$

де $m(x)$ – детермінована функція, що описує «структурований» компонент z у x (тренд); $\varepsilon'(x)$ – складова, що являє собою локальні стохастичні, але просторово корельовані відхилення від $m(x)$ – регіоналізована змінна, і ε'' – залишок, просторово незалежний гауссівський шум, що має нульове середнє і дисперсію σ^2. Тоді інтерполяція може бути отримана як зважена сума даних:

$$z(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i) , \qquad (6)$$

де $z(x_0)$ – точка, в якій шукається значення;

$z(x_i)$ – значення в i-тій точці;

λ_i – невідома вага для виміряного значення в i-тій точці;

n – кількість опорних точок.

Формула схожа на метод середнього зважування обернено пропорційно відстані, але в даному випадку вага точки залежить від варіограми та просторових взаємозв'язків опорних точок.

Метод кригінгу поділяється на ординарний кригінг, який застосовують у більшості випадків, та універсальний, який передбачає, що в даних є тенденція до домінування окремих значень. Універсальний кригінг використовують у випадках, коли відомо, що дані містять науково підтверджені тенденції.

Кригінг може надавати більш точні результати разом із засобами їх верифікації. У більшості випадків при середньомасштабному моделюванні геохімічних полів доцільно використовувати універсальний кригінг, який дозволяє виділяти та контролювати вплив просторових трендів на поширення показників. Слабкою стороною використання кригінгу є складність оброблення даних та обмеження на роботу з аномальними значеннями.

Важливість геостатистичних методів у географічному моделюванні привела до створення окремого модуля «Geostatistical Analyst» у складі ArcGIS. Модуль містить засоби для побудови і аналізу варіограми, проведення інтерполяції методами кригінгу і кокригінгу.

Аналіз геохімічних даних складається з виявлення загальних і регіональних трендів у розподілі хімічних елементів, середніх значень вмісту елемента, які притаманні фізико-географічним районам, виявлення аномальних значень. Саме геохімічним полям притаманна можливість утворення локальних аномалій, які можуть існувати на обмеженій території. Урахування цих аномалій є особливо важливим при пошуку родовищ і дослідженні впливу токсичних хімічних елементів на здоров'я людини.

Під час моделювання на основі геологічної бази даних було вибрано метод кригінгу (6). Метод дозволяє плавно інтерполювати значення геохімічної поверхні і, разом з тим, відобразити аномально високі та низькі значення у растровій моделі.

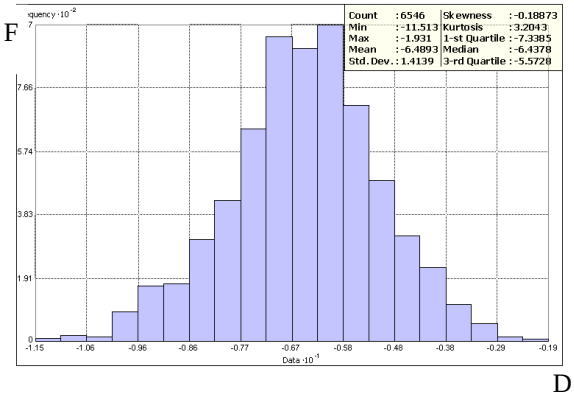

Рис. 2. Гістограма нормалізованого розподілу значень вмісту урану у воді:
F – частоти розподілу, D – значення вимірів

Растрова модель розподілу хімічних елементів є вихідним джерелом для отримання нової інформації. Нові дані можна розподілити на дві основні групи: нові растрові поверхні, нові векторні шари просторової інформації. Нові растрові поверхні можуть бути отримані шляхом перетворення отриманих растрових моделей методами перекласифікації, операцій растрової алгебри. Векторні шари інформації можуть бути отримані як похідні від растрової поверхні. Основними видами векторних даних є ізолінії та полігональні області, що об'єднують точки з однаковими значеннями або діапазонами значень. Залежно від задач моделювання геохімічних полів обирають растрову чи векторну форму подання інформації.

При підготовці електронних карт доцільніше використовувати векторне подання інформації, яке дозволяє зменшити потребу у комп'ютерних ресурсах і забезпечити доступ до атрибутивних значень. За цих умов полігональні векторні дані дозволяють відтворювати геохімічні поля на карті за допомогою класифікації із використанням кількісного фону (рис. 3, 4).

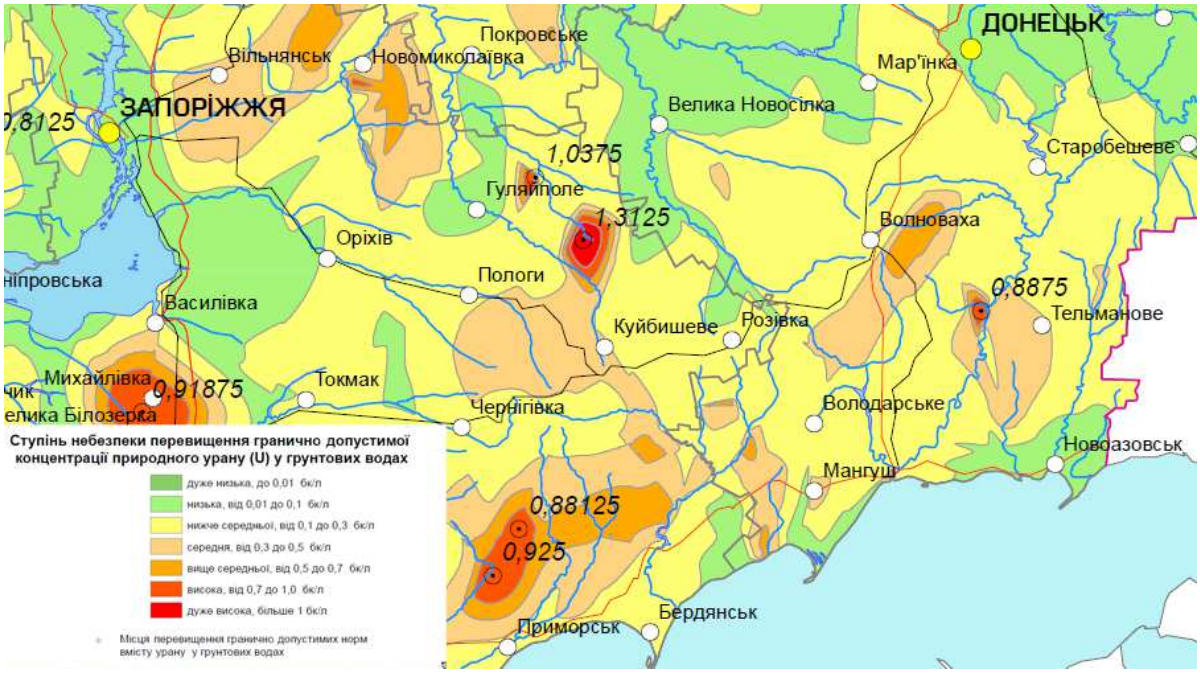

Рис. 3. Фрагмент карти «Небезпека перевищення ГДК природного урану у підземних водах»

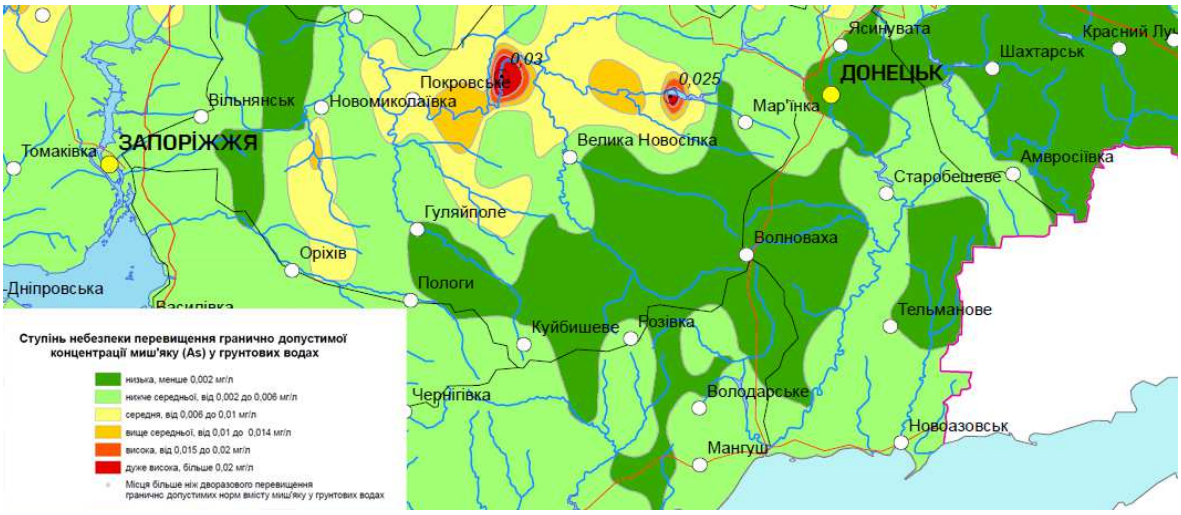

Рис. 4. Фрагмент карти «Небезпека перевищення ГДК миш'яку у підземних водах»

Побудова полігональних векторних об'єктів відбувається у три стадії. На першій стадії проводиться автоматизована побудова полігональних зон через заданий інтервал значень. На другій стадії відбувається узагальнення отриманих полігонів на основі розробленої класифікаційної шкали методом їх автоматизованого об'єднання. У даному випадку за основу класифікації були взяті концентрації хімічних елементів відносно граничних норм вмісту у воді, що дозволило визначити потенційні загрози її вживання для населення. На третьому етапі відбувається генералізація отриманих полігонів з метою запобігання утворенню «паразитарних» полігонів, уникненню випадкових відхилень від регіональних трендів і ліквідації найдрібніших полігонів. Ця робота є вимушеним кроком при роботі з автоматизованими та напівавтоматизованими методами оброблення даних, і її результат залежить від досвіду та знань експерта. На всіх стадіях існує необхідність у контролі топології полігонів.

Для готових тематичних шарів обирають базовий масштаб електронної карти або масштаб паперових карт. Вибір масштабу залежить від вхідних даних і роздільної здатності вихідної растрової інформації згідно з формулою [9]

$$M = \sqrt{\frac{n}{g}}, \qquad (7)$$

де M – іменований масштаб, який визначається числом кілометрів в 1 см;

n – об'єктивне навантаження карти локалізованими об'єктами (кількість об'єктів на 100 см^2);

g – кількість об'єктів на 100 км^2 місцевості.

За допомогою цієї формули та результату векторизації растра вибирають діапазон масштабів, у якому будуть укладені карти. Відповідно до масштабного діапазону вибирають базові географічні шари. Вид і кількість шарів визначаються метою і об'єктом моделювання геохімічних даних. Визначальними при цьому можуть бути елементи гідрологічної мережі, одиниці ландшафтного районування, типи ґрунтів, населені пункти, адміністративні кордони.

Залежно від форми подання карти (електронна чи паперова) наступними стадіями є експорт карти у графічні векторні редактори або обмінний формат електронних карт. У першому випадку основним критерієм є підготовка якісного

поліграфічного макету, у другому – підготовка електронної карти для кінцевих користувачів. При цьому треба підготувати атрибутивні дані, які будуть передані для користування. Розробка формату GeoPDF дозволила публікувати електронні карти у звичному для поліграфічних карт масштабі зі збереженням координатного простору, атрибутивних даних і пошарового відображення.

Висновки

З провадженням геоінформаційних технологій автоматизований інтелектуальний аналіз геохімічних даних став важливим дослідницьким інструментом, який дозволяє з мінімальними витратами часу отримати результати геопросторового розподілу хімічних елементів та оцінити їх за допомогою математичних методів.

Процес створення геостатистичних моделей поєднує підготовку даних на основі їх трансформації та стандартизації, геокодування, побудову геостатистичної моделі, створення кінцевої картографічної продукції та її публікацію.

Основними технічними процесами під час аналізу розподілу хімічних елементів є інтерполяція дискретних точкових даних та їх перетворення у континуальні поверхні, створення векторних ізолінійних шарів. Проведення інтерполяції даних потребує підбору методів та їх параметрів, які найбільш точно відповідають поставленим задачам. Перспективним напрямом є удосконалення моделювання розподілу геохімічних даних на основі геостатистичних методів, у тому числі кригінгу та кокригінгу.

Подальші дослідження пов’язані з удосконаленням методів розроблення геостатистичних моделей геохімічних даних, створенням систем автоматизованого картографування та систем підтримки прийняття рішень на основі ГІС для моніторингу просторового розподілу хімічних елементів та антропогенного забруднення.

Список літератури

1. Згуровский, М. З. Интеллектуальный анализ и системное согласование научных данных в междисциплинарных исследованях / М. З. Згуровский, А. О. Болдак, К. В. Єфремов // Кибернетика и системный анализ. - 2013. – № 4. – С. 62 – 75.

2. Згуровский, М. З. Системный анализ: Проблемы, методология, приложения / М. З. Згуровский, Н. Д. Панкратова. – К.: Наук. думка, 2005. – 743 с.

3. Isaaks, E. H. An Introduction to Applied Geostatistics / E. H. Isaaks, R. M. Srivastava. – Oxford: Oxford Univ. Press, 1989. – 592 p.

4. Clark, H. W. A Practical Geostatistics 2000 / H. W Clark // Publ. by Geostokos Ecosse. – 2004. – 440 p.

5. Wackernagel, H. Multivariate Geostatistics. – [S. l.]: Springer, 2003. – 403 p.

6. Webster, R. Geostatistics for Environmental Scientists / R. Webster, M. O. Oliver. – John Wiley & Sons, 2000. – 286 p.

7. Жуков, В. Т. Компьютерное геоэкологическое картографирование / В. Т. Жуков, Б. А. Новаковский, А. Н. Чумаченко. – М.: Научный мир, 1999. – 128 с.

8. Линник, В. Г. Ландшафтная дифференциация техногенных радионуклидов: геоинформационные системы и модели: автореф. дис. д-ра

геогр.наук: 25.00.23 / РАН, Институт геохимии и аналитической химии им. В. И. Вернадского. – М., 2008. – 40 с.

9. Мусин, О. Р. Цифровые модели «рельефа» континуальных и дискретных географических полей / О. Р. Мусин, С. Н. Сербенюк // Банки географических данных для тематического картографирования. – М.: изд-во Моск. ун-та, – 1987. – С.156 – 170.

10. Перельман, А. И. Геохимия / А. И. Перельман. – М.: Высш. шк., 1975. – 528 с.

11. ArcGIS Geostatistical Analyst: руководство пользователя. – М., 2001. – 219 с.

12. Світличний, О. О. Основи геоінформатики: навчальний посібник / О. О. Світличний , С. В. Плотницький / за заг. ред. О. О. Світличного. – Суми: ВТД «Університетська книга», 2006. – 295 с.

Поступила в редакцию 14.09.2015

Геостатистический интеллектуальный анализ геохимических данных

Рассмотрена методика геостатистического интеллектуального анализа геохимических данных, которая позволяет в автоматизированном режиме выполнить анализ геопространственных закономерностей распространения химических элементов. Определены основные этапы проведения геостатистического анализа данных и подходы к континуальному представлению данных средствами векторных и растровых моделей. Выполнен сравнительный анализ преимуществ и недостатков интерполяции геохимических данных различными методами.

***Ключевые слова**:* интеллектуальный анализ данных, ГИС, геостатистика, интерполяция, геохимия.

Geostatistical intellectual analysis of geochemical data

The method of geostatistical spatial mining of geochemical data is considered, which allows automated analysis of geospatial patterns of chemical elements spread. The main stages of data analysis and geostatistical approaches to continual reporting data of means vector and raster models are detected. A comparative analysis of advantages and disadvantages of the geochemical data interpolation by various methods is held.

Keywords: spatial data mining, GIS, geostatistics, interpolation, geochemistry.