УДК 519.254

М. С. Павлов[1], С. В. Антоненко[1], І. О. Доровська[2]

[1]*Дніпровський національний університет імені Олеся Гончара*

[2]*Криворізька філія ПВНЗ «Європейський університет»*

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОПТИМІЗАЦІЇ МЕРЕЖІ МОНІТОРИНГУ СТАНУ ПІДЗЕМНИХ ВОД З ВИКОРИСТАННЯМ ГЕНЕТИЧНОГО АЛГОРИТМУ

Розглянуто обчислювальну схему оптимізації мережі моніторингу стану підземних вод з використанням генетичного алгоритму на основі крігінгу та методу обернено зважених відстаней.

Ключові слова: *генетичний алгоритм; крігінг; метод обернено зважених відстаней; оптимізація мережі моніторингу.*

Рассмотрена вычислительная схема оптимизации сети мониторинга состояния подземных вод с использованием генетического алгоритма на основе кригинга и метода обратных взвешенных расстояний.

Ключевые слова: *генетический алгоритм; кригинг; метод обратных взвешенных расстояний; оптимизация сети мониторинга.*

Groundwater level monitoring networks can provide significant information about water resources in certain area. Cost effective network design is very important for the management. Maintenance cost of the existing monitoring network can be reduced by minimizing the number of the monitoring wells which do not provide useful information or this information can be reproduced. Finding the group of wells, where measurements could be estimated with the lowest error may be a combinatorial problem as there are many possible combinations to check. Genetic algorithm has a great potential for solving such problems because the time spent searching for a satisfactory solution may be less than the time spent for an exhaustive search, but it does not guarantee the optimal result. In this paper we consider the computational scheme for optimizing existing groundwater monitoring network located in Kryvyi Rih, Dnipropetrovsk region, Ukraine, using genetic algorithm based on ordinary kriging technique and inverse distance weighting method. To find the groups of the redundant wells Python application for the network design optimization was developed. The solution was found for the network with 46 monitoring wells. The number of wells to remove was set from 3 to 7. In terminology of the genetic algorithm each possible solution candidate is called a «chromosome». Chromosome is characterized by a set of values known as

© Павлов М. С., Антоненко С. В., Доровська І. О., 2018

«genes». In our case, chromosome is a set of monitoring wells. Iteration of the algorithm is called a «generation». Every generation produced the group of possible solution candidates called a «population». Ordinary kriging technique and inverse distance weighting method were used to estimate the measurements for the candidates. Based on actual and estimated values root-mean-square error for each candidate was calculated. Selection was applied to each population to find the candidates with the smallest error and pass them to the next generation. Single-point crossover was used to exchange genes among chromosomes. To maintain diversity within the population procedure known as «mutation» was applied. Mutation occurred for each gene in the chromosome with the 10% probability. In case of mutation, the gene was replaced by a random one not already included in the chromosome. The number of generations was predefined. Based on the results further development of the tool for the groundwater level monitoring network optimization is planned.

Keywords: *genetic algorithm; kriging; idw; groundwater monitoring network.*

Постановка проблеми в загальному вигляді. Довгострокові мережі моніторингу підземних вод можуть надавати важливу інформацію для планування та управління водними ресурсами. Бюджетні обмеження в органах управління водними ресурсами часто призводять до зменшення кількості спостережних свердловин, включених до мережі моніторингу. Дані з цих мереж можуть бути використані для перевірки моделей підземних вод, перегляду правил щодо довгострокової стійкості водних ресурсів, для оцінки реакції рівнів ґрунтових вод на зусилля з поповнення та зміни кліматичних умов. Враховуючи високі витрати, пов'язані з підтримкою цих мереж, розробка ефективних мережевих конструкцій є дуже важливою.

Зменшення кількості спостережних свердловин в наявній мережі моніторингу є нелінійною комбінаторною проблемою. Надзвичайно актуальним є питання розробки технологій з використання евристичних процедур оптимізації.

Аналіз останніх досягнень. Для довгострокового моніторингу рівнів води типовою метою є розробка економічно ефективного проекту управління водними ресурсами, що зберігає адекватну загальну точність прогнозування. У роботі [1] запропоновано три моделі оптимізації вибору підмножини станцій з великої мережі моніторингу підземних вод, для вирішення яких використовують алгоритм імітації відпалу. В роботі [2] для оптимізації мережі моніторингу застосовували алгоритм оптимізації мурашиної колонії.

Мета роботи. Метою роботи є розробка нової обчислювальної технології, яка дозволяє на сучасному рівні проводити оптимізацію мереж моніторингу.

Основна частина. Просторова автокореляція кількісно визначає основний принцип географії – речі, які ближче, більш схожі, ніж речі, відокремлені один від одного. Сформулюємо основні теоретичні положення.

Метод обернено зважених відстаней (англ. Inverse Distance Weighting – IDW) передбачає, що об'єкти, які знаходяться поблизу, більш подібні один до одного, ніж об'єкти, віддалені один від одного [3]. Щоб проінтерполювати значення для деякого положення x, IDW використовує виміряні значення навколо x. Найбільш близькі до невідомого значення сильніше впливають на прогнозоване значення, ніж віддалені від нього на значну відстань. Тому метод надає більші ваги точкам, розташованим ближче всього до невідомого значення.

В IDW ваги базуються на відстані d_{ix} від кожної з відомих точок z_i до точки, яку ми намагаємося оцінити z_x, i={1,n}, де x та i – узагальнене позначення координат. В IDW ми розглядаємо обернену відстань $\frac{1}{d_{ix}}$. Тобто z_i отримує значення ваги w_i за формулою (1):

$$w_i = 1/d_{ix} \frac{1}{\sum_{i=1}^{n} 1/d_{ix}} \ , \ \sum_{i=1}^{n} (w_i) = 1 \qquad (1)$$

де z_i – виміряне значення в i-ій локації;

w_i – вага, присвоєна z_i відносно відстані до z_x.

Невідоме значення в локації z_x отримується наступним чином:

$$z(x) = \sum_{i=1}^{n} (w_i z_i) \qquad (2)$$

Особливість звичайного крігінга (англ. ordinary kriging) полягає в тому, що він дозволяє оцінити похибку кожного прогнозування, забезпечуючи міру впевненості в моделюваній поверхні, і з цієї причини вважається статистичною технікою, а не детерміністичним методом. Крігінг схожий на IDW, оскільки він обтяжує навколишні виміряні значення, щоб отримати прогноз для відомого місця. В IDW вага w_i залежить виключно від відстані до місця прогнозування. Проте в крігінгу ваги базуються не тільки на відстані між виміряними точками та місцем прогнозування, а й на загальному просторовому розташуванні цих точок. Щоб використовувати просторове розташування при обчисленні ваги, необхідно проаналізувати просторову автокореляцію. Оцінка звичайним крігінгом може бути виражена як:

$$z(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i), \ \sum_{i=1}^{n} \lambda_i = 1 \qquad (3)$$

де $z(x_0)$ – невідоме значення в локації x_0;

$z(x_i)$ – виміряне значення в x_i локації;

λ_i – ваговий коефіцієнт для виміряного значення в i локації;

x_0 – місце прогнозування;

n – кількість відомих точок.

Для визначення вагових коефіцієнтів λ_i, що забезпечують мінімальну похибку для заданого набору просторових даних використовується варіограмна модель.

Крігінг використовує властивість, що називається *семіваріація*, для вираження ступеня взаємозв'язку між точками на поверхні. Семіваріація – це лише половина дисперсії різниці між усіма можливими точками, розташованими на відстані h. Семіваріація на відстані $h=0$ буде нульовою, оскільки між точками, які порівнюються з собою, немає різниці. Проте, коли точки порівнюються з більш віддаленими точками, семіваріація зростає. На деякій відстані, яка називається *діапазоном*, семіваріація стане приблизно рівною дисперсії всієї поверхні. Це найбільша відстань, через яку величина в точці поверхні пов'язана зі значенням в іншій точці. Діапазон визначає максимальне сусідство, в якому слід вибирати контрольні точки для оцінки вузла сітки поверхні, щоб скористатися статистичною кореляцією серед спостережень.

Графік залежності між значеннями відстані та семі варіації, побудований на основі вибіркових даних, називають експериментальною варіограмою. Зазвичай є багато пар точок, кожна пара має унікальну відстань. Побудова на графіку експериментальної варіограми з усіх пар швидко стає незручною. Замість того, щоб будувати кожну пару, пари групують відповідно до відстані, тобто за h береться не конкретна відстань, а деякий діапазон або крок. Значення варіограми розраховується наступним чином:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)|h_{ij}=h} (x_i - x_j)^2 \qquad (4)$$

Значення $\gamma(h)$ при $h = 0$, яке ще позначається як c_0 – це залишкова варіація, тобто дисперсія похибок вимірювань [4]. Зі збільшенням відстані варіограма зростає до максимальних значень при деякому значенні a, яке називають *радіусом кореляції*. При подальшому збільшенні відстані варіограма не збільшується, тобто втрачається залежність різниці значень в точках від відстані між ними. Це називається *порогом* варіограми. Приклад варіограми зображений на рис. 1.

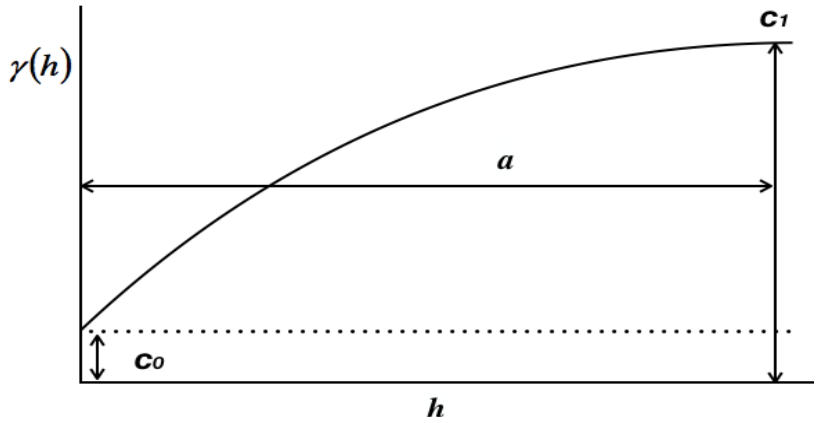

Рисунок 1 – Експериментальна варіограма з позначенням основних параметрів

Для обчислення значень варіограми для невідомих точок підбирається модель варіограми. Форма експериментальної варіограми допомагає правильно підібрати цю модель та її параметри.

Найбільш широко на практиці використовують такі моделі:

сферична модель

$$\gamma(h) = c_0 + c_1[1.5\left(\frac{h}{a}\right) - 0.5\left(\frac{h}{a}\right)^3], h <= a; \quad (5)$$
$$\gamma(h) = c_0 + c_1, h > a;$$

експоненціальна модель

$$\gamma(h) = c_0 + c_1[1 - exp\left(-\frac{h}{a}\right)] \quad (6)$$

гауссівська модель

$$\gamma(h) = c_0 + c_1[1 - exp\left(-\frac{h^2}{a^2}\right)] \quad (7)$$

де h – крок(відстань), a – радіус кореляції, c_1 – порогове значення варіограми, c_0 – залишкова варіація.

Отже, дисперсія оцінки змінної $z(x_0)$ може бути записана як функція значень варіограми між усіма парами точок x_i x_j, значень варіограми між усіма x_i та оцінюваною точкою x_0 та значень вагових коефіцієнтів:

$$\sigma^2 = \sum_{i=1}^{n} \lambda_i \gamma(x_i, x_0) + \phi \quad (8)$$

де σ^2 – дисперсія відхилень прогнозованого значення змінної в точці оцінювання від істинного.

Тобто, мінімальна похибка отримується коли:

$$\sum_{i=1}^{n} \lambda_i \gamma(x_i, x_j) + \phi = \gamma(x_i, x_0), \text{ для всіх } j \qquad (9)$$

де $\gamma(x_i, x_0)$ – семіваріація на відстані від x_i до x_j,

$\gamma(x_i, x_0)$ – семіваріація на відстані від x_i до x_0,

ϕ – лагранжіан для вирішення рівняння,

Задача інтерполяції полягає в знаходженні такого набору вагових коефіцієнтів λ_i, який би забезпечував максимальну точність оцінки, тобто мінімальну дисперсію σ^2. Таким чином, у звичайному крігінгу вага λ_i залежить від пристосованої моделі, відстані до місця прогнозування та просторових співвідношень між виміряними значеннями навколо місця прогнозування.

Для оцінювання точності моделі використовується *перехресна перевірка*. Перехресна перевірка вилучає задану кількість свердловин з набору даних (*N*) і оцінює рівень води у видалених місцях шляхом крігінгу (або IDW). Потім розраховується похибка оцінки, тобто різниця між фактичним (*z*) та розрахунковим значенням (z^*) у місці вилученої свердловини (z-z^*). Ця процедура повторюється для кожної унікальної групи в наборі даних, та підраховується *середньоквадратична похибка* (англ. RMSE). Кандидати на вилучення з мережі – це групи свердловин, видалення яких призводить до найменшої середньоквадратичної похибки.

$$RMSE = \sqrt{\frac{1}{N}\sum(z(x_i) - z*(x_i))^2} \qquad (10)$$

Зменшення кількості спостережних свердловин в наявній мережі моніторингу є нелінійною комбінаторною проблемою. Доречним є використання евристичних процедур оптимізації.

Генетичний алгоритм – це адаптивний евристичний алгоритм пошуку, що імітує механіки природного відбору і добре підходить для вирішення задач комбінаторної оптимізації, в яких існує великий набір варіантів рішень.

Впровадження генетичного алгоритму починається з генерації випадкових хромосом. У термінології генетичного алгоритму масив рішень (або рядок генів) в задачі оптимізації називається *хромосомою.* В нашому випадку – це випадкові комбінації номерів свердловин, кандидатів на видалення. Тобто хромосома являє собою унікальне рішення в просторі рішень, множині всіх рішень в оптимізаційній проблемі. Кожна ітерація генетичного алгоритму називається *генерацією*. Під час кожної генерації виконуються наступні дії:

- обчислюється функція *фітнесу* (англ. fitness), тобто для кожного кандидата вираховується значення помилки прогнозування на основі крігінгу (або IDW) та перехресної перевірки.
- проводиться *селекція* (англ. selection), тобто кандидати з мінімальним значення помилки копіюються до нової популяції. Це гарантує, що найкращі рішення можуть вижити до кінця пробігу генетичного алгоритму.
- проводиться *схрещування* хромосом (англ. crossover): найкращі кандидати породжують нащадків. Тобто одна частина хромосоми копіюється від першого з батьків (приблизно половина), а друга – від другого. Існує шанс представити дубльовані свердловини у нащадків під час схрещування. Щоб запобігти цьому, нащадок з дублікатами видаляється.
- проводиться процес мутації кожної хромосоми. Гени (тобто номери свердловин) заміщуються (в нашому випадку з 10 % вірогідністю) на інші з вибірки. Якщо мутація призводить до дублікатів в хромосомі, процес мутації гену повторюється. Мутація гену відхиляється, якщо за декількох спроб не обрано жодного гену, який би не призводив до дублікатів в хромосомі.
- процес повторюється, доки не відбулася встановлена кількість генерацій або не було знайдено кандидата з прийнятним (заданим) рівнем помилки прогнозування.

За допомогою можливостей бібліотек мови програмування Python розроблено програмний модуль, який дозволив перенести вже накопичені дані з файлів формату .DBF до створеної бази даних. Розроблено інформаційну систему, яка дозволяє зберігати та обробляти дані, працювати з файлами різного формату, проводити статистичний аналіз вибірок, оптимізацію мереж, відображати результати досліджень у вигляді таблиць та графіків. Використання виконується через веб-інтерфейс. Структурну схему програмної системи приведено на рис. 2.

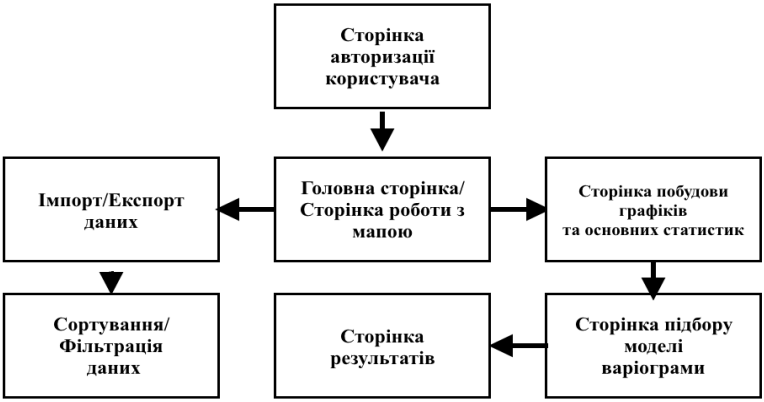

Рисунок 2 – Структурна схема програмної системи

Розглянемо вирішення поставленої задачі щодо оптимізації мережі моніторингу, тобто пошуку кандидатів на вилучення з мережі, на прикладі мережі із 46 свердловин. Для тестування алгоритмів було використано набір даних рівнів води зі свердловин біля хвостосховища у м. Кривий Ріг, Україна. Джерело інформації про спостережні свердловини включає географічні координати (тобто довгота, широта) і значення рівня води відповідної свердловини. Дані моніторингу збиралися за календарний рік, спостереження проводились 6, 12, 18, 24, 30 числа кожного місяця. Для аналізу використовувалися медіанні значення відстані до води для кожної свердловини. Розташування свердловин зображено на рис. 3.

Основні характеристики вибірки з даних приведені у табл. 1.

Побудовану гістограму зображено на рис. 4.

З характеристик гістограми (рис 4.) випливає, що ряд однорідний і має відносно невеликий розкид, про що свідчить коефіцієнт варіації:

$$V = \sigma / x,$$

де σ–стандартне відхилення, x – середня відстань. Статистичні значення вибірки приведені у табл. 2.

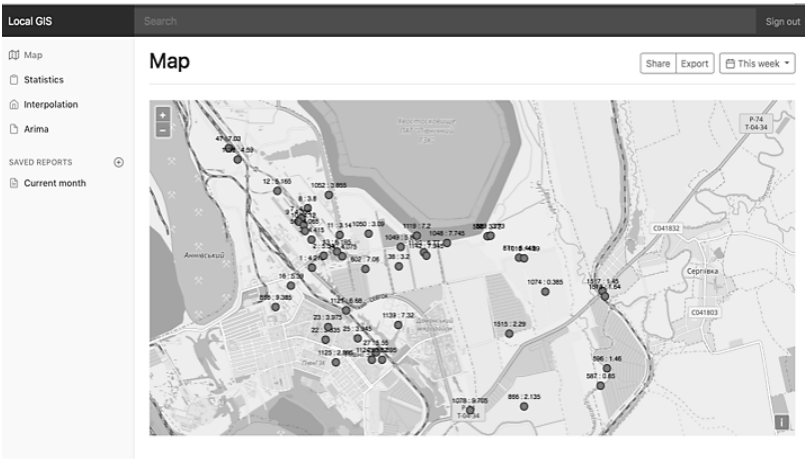

Рис.3 – Розташування свердловин

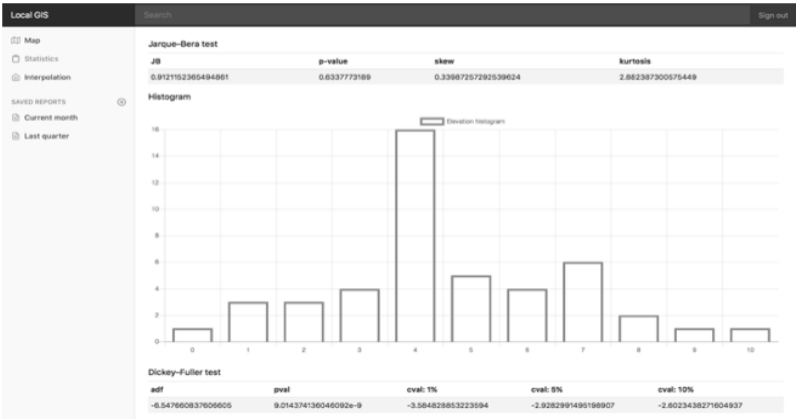

Рисунок 4 – Гістограма

Таблиця 1

Базові характеристики вибірки

Кількість свердловин	Мінімальна відстань до води (м)	Максимальна відстань до води (м)	Середня відстань до води (м)
46	0,385	9,705	4,6

Таблиця 2

Базові статистики вибірки

Мода	Стандартне відхилення	Дисперсія	Коефіцієнт варіації
4,1	2,11	4,46	0,4587

Далі побудовано експериментальну варіограму (рис. 5), на її основі підібрано модель варіограми та параметри. Було вирішено використовувати експоненціальну модель варіограми з параметрами $c_0 = 0; a = 7{,}97; c_1 = 5089{,}18$.

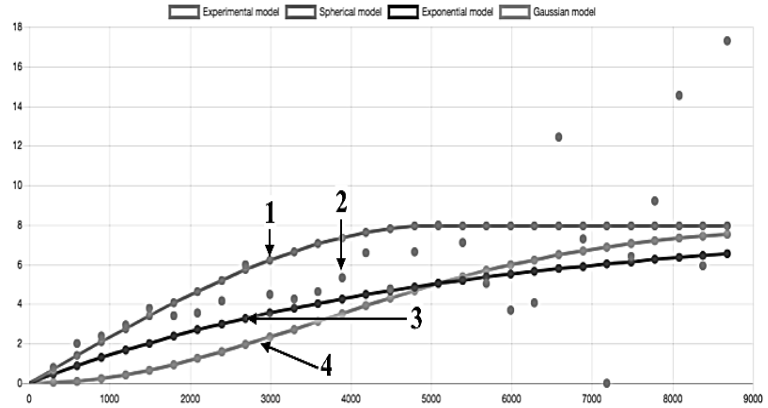

Рисунок 5 – Моделі варіограми

На рис. 5 представлено моделі варіограми: 1 – сферична модель; 2 – експериментальна модель; 3 – експоненціальна модель, 4 – гауссівська модель.

У табл. 4 та табл. 5 приведено результати роботи алгоритму в разі повного перебору можливих рішень з використання IDW та відповідно – крігінгу. Приведено приблизний час розрахунку відносно шуканої кількості свердловин на видалення з мережі. Обрахунок усіх комбінацій (15180) із трьох свердловин, видалення та прогнозування на їхніх місцях значень за допомогою крігінгу або IDW у цьому випадку займає приблизно годину. Для групи з чотирьох свердловин час зростає до 12 годин для 163185 комбінацій, для групи з 5-ти вже більше 100 годин при 1370754 комбінаціях.

Як видно, зменшення кількості спостережних свердловин в наявній мережі моніторингу є нелінійною комбінаторною задачею, з огляду на представлену кількість унікальних рішень (табл. 4, 5). Зважаючи на це, повним перебором розраховувалась лише група з 3-ох свердловин. З використанням IDW було знайдено рішення – свердловини № 3, № 7, № 8 зі значенням RMSE=0,13, а найкраще рішення для видалення – свердловини № 1, № 1052, № 1121 із найменшим значенням середньоквадратичної похибки – 0,03, знайдене з використанням крігінгу.

Завдяки генетичному алгоритму витрати часу на пошук задовільного рішення можна звести до мінімуму. Результати роботи генетичного алгоритму представлені у табл. 6 та у табл. 7. Для пошуку задовільного рішення для групи з 3-ох свердловин знадобилось 1,2 хвилини, використовуючи метод обернено зважених відстаней, при цьому значення RMSE=0,18, номери свердловин на видалення відповідно № 3, № 9, № 1016. З використанням крігінгу знадобилось 2,4 хвилини, свердловини № 3, № 50, № 25, зі значенням RMSE=0,21. Це гарний приклад того, що алгоритм не гарантує знаходження найкращого рішення, але переважає у часі пошуку, що надає можливість шукати допустимі рішення у великих мережах за прийнятний час. Хоча, наприклад, для видалення групи з 4-ох свердловин з мережі з використанням крігінгу було знайдено рішення, близьке до оптимального (свердловини № 588, № 50, № 12, № 602 відповідно, зі значенням RMSE=0,07), що є гарним результатом.

Таблиця 4

Результати роботи при повному переборі з використанням IDW

Кількість	**Унікальних**	**t (год) повний перебір**	**IDW (повний перебір)**	**IDW RMSE**
3	15180	1,1	3, 7, 8	0,13
4	163185	12,1	–	–
5	1370754	101,5	–	–
6	9366819	693,6	–	–
7	53524680	3963,2	–	–

Таблиця 5

Результати роботи при повному переборі з використанням крігінгу

Кількість	**Унікальних**	**t (год) повний перебір**	**Крігінг (повний перебір)**	**Крігінг RMSE**
3	15180	1,1	1, 1052, 1121	0,03
4	163185	12,1	–	–
5	1370754	101,5	–	–
6	9366819	693,6	–	–
7	53524680	3963,2	–	–

Таблиця 6

Результати роботи з використанням генетичного алгоритму та IDW

Кількість	**t (год)**	**Кількість генерацій**	**IDW (GA) № свердловин**	**IDW RMSE**
3	0,02	10	3, 9, 1016	0,18
4	0,05	14	1052, 1138, 8, 9	0,2
5	0,03	7	1, 1138, 1052, 50, 8	0,15
6	0,05	9	25, 1124, 23, 28, 3,7	0,13
7	0,06	11	25, 1138, 14, 1, 22, 28, 1124	0,13

Таблиця 7

Результати роботи з використанням генетичного алгоритму та крігінгу

Кількість	**t (год)**	**Кількість генерацій**	**Крігінг (GA) № свердловин**	**Крігінг RMSE**
3	0.04	20	3, 50, 25	0,21
4	0.09	21	588, 50, 12, 602	0,07

Продовження таблиці 7

Кількість	**t (год)**	**Кількість генерацій**	**Крігінг (GA) № свердловин**	**Крігінг RMSE**
5	0.18	32	12, 7, 50, 1515, 25	0.27
6	0.23	35	27, 3, 50, 22, 602, 866	0.22
7	0.31	39	10, 866, 27, 9, 602, 28, 596	0.41

Завдяки випадковості, яка закладена в генетичний алгоритм, кожен перезапуск його дає можливість покращити попередній результат.

Висновки та перспективи подальшого розвитку. Створена в цій роботі нова обчислювальна технологія з використанням генетичного алгоритму на базі крігінгу та методу обернено зважених відстаней дозволяє виконувати оптимізацію мереж гідрогеологічного моніторингу. Це сприятиме зниженню ціни на обслуговування вже наявних мереж та удосконаленню процесу розташування нових. Слід зазначити, що евристичні алгоритми пошуку мають великий потенціал у вирішенні подібних комбінаторних проблем. На підставі отриманих результатів планується подальша розробка інтелектуальної системи оптимізації мережі гідрогеологічного моніторингу.

Бібліографічні посилання

1. Nunes L. M., Cunha M. C., Ribeiro L. Groundwater monitoring network optimization with redundancy reduction. // Journal of Water Resources Planning and Management. V. 130, No. 1. 2004. P. 33–43.

2. Li Yuanhai, Hilton A. B. C. Reducing Spatial Sampling in Long-Term Groundwater Monitoring Networks Using Ant Colony Optimization. International Journal of Computational Intelligence Research, V. 1, No. 1, 2005, p. 19–28.

3. Шипулін В. Д., Основи ГІС-аналізу: навч. посібник Харк. нац. ун-т міськ. госп-ва ім. О. М. Бекетова. Х. : ХНУМГ, 2014. 264 с.

4. Світличний О. О., Плотницький С. В. Основи геоінформатики: Навчальний посібник. Суми: ВТД «Університетська книга», 2006. 201 с.

Надійшла до редколегії 11.11.2018 р.