

КОРЕЛЯЦІЙНО-РЕГРЕСІЙНИЙ АНАЛІЗ КОНЦЕНТРАЦІЇ ЗАБРУДНЕННЯ В ҐРУНТОВОМУ РОЗЧИНІ ІНФІЛЬТРАЦІЙНОГО ШАРУ

А. О. Курило, аспірант

Сумський національний аграрний університет

В статті за допомогою парного кореляційно-регресійного аналізу встановлено взаємозв'язок концентрації забруднення в ґрунтовому розчині інфільтраційного шару та часу випадіння опадів. Встановлено форму залежності між досліджуваними змінними, оцінено функцію регресії та здійснено прогноз залежної змінної.

Ключові слова: концентрація забруднення, час випадіння опадів, парний кореляційно-регресійний аналіз, прогноз.

Постановка проблеми. За останні десятиліття нагромадження та міграція рухомих форм техногенних забруднень на сільськогосподарських землях набули загрозливих розмірів. Така міграція відбувається в процесі несистематичного природного зволоження інфільтраційного шару ґрунту.

У зв'язку з цим, постає потреба систематично наперед оцінювати стан екологічної безпеки та прогнозувати тенденції характерних показників стану екосистеми. Це дасть змогу приймати оперативні рішення щодо безпеки навколишнього середовища і рекомендувати ефективні оперативні заходи із захисту населення та екосистем.

Останні дослідження та публікації. Серед науковців все частішими стають дослідження питань застосування математичних методів в дослідженнях закономірностей протікання явищ та процесів при оцінці стану екологічної безпеки. Зокрема вивчення та прогнозування гідрогеологічних процесів методами математичного моделювання описані авторами Євграшкіною Г.П., Шерстюк Н.П. [1, 2]. Бойко Т.В в своїй роботі розглядає використання математичних моделей в сфері гідро-екології, що дає змогу проаналізувати вплив техногенного навантаження на ґрунти [3]. Для спрощення системи автор вдається до зменшення розмірності і пропонує двовимірну модель міграції забруднюючих величин у ґрунтах.

Моніторинг родючості ґрунтів вимагає проведення досліджень на високому науковому рівні, з застосуванням сучасного аналітичного обладнання з відповідним комп'ютерним забезпеченням, здатним оперативно математично обробляти та узагальнювати отриману аналітичну інформацію [4]. Для прогнозування екологічного стану довкілля на регіональному рівні в роботі [5] запропоновано метод експертних оцінок у комплексі з багатфакторним аналізом.

Тому, вважаючи одним із найважливіших застосування методів математичної статистики для дослідження взаємозв'язку двох величин, у статті більш детально зосередимо увагу на базовій темі цього напрямку – парному кореляційно-регресійному аналізі.

Формулювання цілей статті. Основною метою дослідження є встановлення залежності концентрації забруднення в ґрунтовому розчині інфільтраційного шару від часу випадіння опадів. На основі статистичних даних, отриманих по 30-ти замірах в п'яти різних точках на лінії річки, виконуємо розрахунки засобами парного кореляційно-регресійного аналізу, вбудованими в Microsoft Excel 2010. Завданнями саме регресійного аналізу є встановлення форми залежності між досліджуваними змінними, оцінка функції регресії та прогноз значень залежної змінної. В табл. 1 наведені статистичні дані концентрації забруднення та часу випадіння опадів для першої контрольної точки. Незалежною змінною x буде час випадіння опадів, а залежною змінною $y = -Ln \frac{C}{C_p}$ –

логарифм відношення концентрації забруднення в ґрунтовому розчині інфільтраційного шару до початкової концентрації забруднення в застійних зонах ґрунту (надалі – логарифм відносної концентрації забруднення) [6].

Виклад матеріалу дослідження. Парний коефіцієнт кореляції r – міра тісноти лінійного зв'язку двох випадкових величин X та Y , поданих у вигляді двох вибірок однакового обсягу [7]. За знаком та абсолютним значенням коефіцієнта кореляції можна зробити такі висновки:

1. Чим ближче абсолютне значення коефіцієнта кореляції до 1, тим тісніший зв'язок існує між двома змінними x та y , а чим ближче воно до нуля, тим зв'язок слабкіший.

2. Якщо коефіцієнт кореляції позитивний $r > 0$, то зв'язок між змінними прямий, тобто при збільшенні незалежної змінної x залежна змінна y теж збільшується. При $r < 0$ зв'язок між змінними зворотний, тобто при збільшенні незалежної змінної x залежна змінна y зменшується.

Для перетворення кількісної характеристики тісноти лінійного зв'язку між двома випадковими змінними у якісну, застосуємо шкалу англійського статистика Чеддока.

Для проведення статистичних розрахунків у Microsoft Excel 2010 скористаємось надбудовою для аналізу наукових даних, а саме інструментом «Регресія» [8]. Реалізація цих розрахунків у

Microsoft Excel представлена на рисунку 1.

Таблиця 1.

Статистичні дані осередненої концентрації забруднення та інтегрального часу випадання опадів

№ з/п	Інтегральний час випадання опадів τ , години	Логарифм відносної концентрації забруднення	№ з/п	Інтегральний час випадання опадів τ , години	Логарифм відносної концентрації забруднення
1	0	0	16	3,75	0,071
2	0,25	0,009	17	4	0,125
3	0,5	0,018	18	4,25	0,089
4	0,75	0,020	19	4,5	0,097
5	1	0,015	20	4,75	0,084
6	1,25	0,022	21	5	0,097
7	1,5	0,032	22	5,25	0,084
8	1,75	0,039	23	5,5	0,092
9	2	0,034	24	5,75	0,125
10	2,25	0,043	25	6	0,114
11	2,5	0,046	26	6,25	0,149
12	2,75	0,071	27	6,5	0,128
13	3	0,060	28	6,75	0,125
14	3,25	0,067	29	7	0,155
15	3,5	0,068	30	7,25	0,190

Джерело: статистичні дані, зібрані автором.

	R	S	T	U	V	W	X
23	ВЫВОД ИТОГОВ						
24							
25	Регрессионная статистика						
26	Множественный R	0,953690405					
27	R-квадрат	0,909525389					
28	Нормированный R-квадрат	0,906294153					
29	Стандартная ошибка	0,014849141					
30	Наблюдения	30					
31							
32	Дисперсионный анализ						
33		df	SS	MS	F	Значимость F	
34	Регрессия	1	0,062065289	0,062065289	281,47909	3,84611E-16	
35	Остаток	28	0,006173915	0,000220497			
36	Итого	29	0,068239204				
37							
38		Коэффициенты	Стандартная ошибка	t-статистика	P- Значение	Нижние 90,0%	Верхние 90,0%
39	Y-пересечение	-0,000685527	0,005289332	-0,129605528	0,8978051	-0,009683373	0,008312
40	Переменная X 1	0,021020088	0,001252886	16,77733853	3,846E-16	0,018888765	0,023151

Рис. 1. Результати розрахунків інструментом Регресія в Microsoft Excel

Отже, так як парний коефіцієнт кореляції між логарифмом відносної концентрації забруднення та часом випадіння опадів дорівнює $r=0,954$, то за шкалою Чеддока між цими величинами існує дуже сильний лінійний зв'язок, причому цей зв'язок прямий.

Оскільки парний коефіцієнт кореляції обчислюється для випадкових величин, то він теж є випадковою величиною. Тому для нього будемо висувати і перевіряти гіпотезу про те, чи статистично значуще він відрізняється від нуля. Цю гіпотезу перевіряємо за допомогою t -критерію Стюдента. Отримуємо фактичне значення критерію Стюдента $t_{\text{факт}} = 16,8$ та порівнюємо його із критичним $t_{0,1;28}=1,7$ для степенів свободи $k = n - 2 = 28$ й вірогідності $p = 0,9$ (можна обирати

довільно), визначеним за допомогою функції СТЬЮДРАСПОБР Microsoft Excel.

Оскільки фактичне значення критерію Стюдента більше ніж табличне ($16,8 > 1,7$), то парний коефіцієнт кореляції статистично значуще відрізняється від 0 при вірогідності 0,9 й залежність між логарифмом відносної концентрації забруднення та часом випадіння опадів є достовірною.

Парна лінійна регресія – причинна модель статистичного лінійного зв'язку між двома кількісними змінними Y та X , подана рівнянням:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon, \quad (1)$$

де β_0, β_1 – параметри моделі; ε – випадкова величина, що включає вплив не врахованих у

моделі факторів, випадкових похибок та особливостей вимірювання, викликаних специфікацією моделі, вибірковою характеристикою вихідних даних та особливостями вимірювання змінних [7].

В основі оцінки рівняння парної лінійної регресії $\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$ лежить вибірка 30 пар значень незалежної та залежної змінних. Геометричною інтерпретацією такої вибірки буде сукупність 30 точок на площині, яка називається діаграмою розсіювання або кореляційним полем (рис. 2).

Задачею парного лінійного регресійного

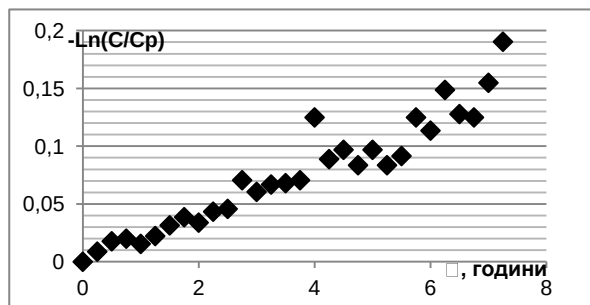


Рис. 2. Геометрична інтерпретація вибірок статистичних даних – кореляційне поле

Згідно з ним, оцінки невідомих параметрів β_0 , β_1 обираються таким чином, щоб сума квадратів відхилень теоретичних значень, знайдених за рівнянням $\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$, від емпіричних значень y_i , була мінімальною [7, 9].

Далі, виконання необхідної умови екстремуму функції двох змінних β_0 та β_1 уможливорює їх оцінку за простими формулами:

$$\beta_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}, \quad (2)$$

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}, \quad (3)$$

де $\bar{x}$, $\bar{y}$ – середні значення x та y , $\overline{xy}$ – середній добуток, $\overline{x^2}$ – середній квадрат.

По вихідних статистичних даних (див. табл. 1) оцінюємо засобами Microsoft Excel параметри регресії (рис. 1): $\beta_1 = 0,0210$, $\beta_0 = -0,0007$.

Модель парної лінійної регресійної залежності логарифма відносної концентрації забруднення від часу випадіння опадів має вигляд:

$$Y = -0,0007 + 0,0210 \cdot X + \varepsilon. \quad (4)$$

Параметр $\beta_0 = -0,0007$ не має жодного змісту, так як він від'ємний. $\beta_1 = 0,021$ показує оцінку зміни логарифма відносної концентрації забруднення з кожною годиною випадіння опадів без урахування випадкових та неврахованих у моделі чинників.

Параметри регресії оцінені нами на основі

аналізу в нашому дослідженні є знаходження такої прямої, яка найкращим чином аналітично характеризує кореляційне поле, зображене на рис. 2. Тобто потрібно мінімізувати деяку «міру» відхилень ε_i від лінії регресії, представленої на рис. 3.

Найбільш поширеною мірою відхилення є сума квадратів відхилень. Метод знаходження оцінок параметрів регресії за допомогою цієї функції називається методом найменших квадратів. Його алгоритм реалізований багатьма комп'ютерними програмами, зокрема в Microsoft Excel – інструментом «Регресія».

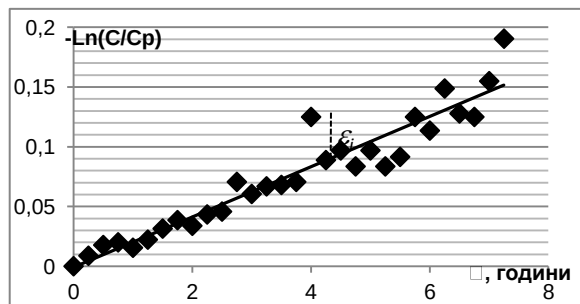


Рис. 3. Геометрична інтерпретація вибірок статистичних даних – кореляційне поле з лінією регресії

вибірки емпіричних величин, в значеннях яких значний внесок зроблено випадковими факторами. Тому в оцінці кожного із параметрів наявна суттєва випадкова складова. Через неї оцінені параметри потребують перевірки статистичної значущості та побудови довірчих інтервалів. З цією метою визначимо стандартну помилку за кожним з параметрів – m_{β_0} та m_{β_1} . [7, 9] за відомими формулами або ж інструментом «Регресія» (рис. 1): $m_{\beta_0} = 0,0053$, $m_{\beta_1} = 0,0013$.

Стандартну помилку регресії можна знайти за формулою або ж інструментом «Регресія» (рис. 1): $S = 0,0149$.

Фактичні значення t -критерію Стюдента для перевірки статистичної значущості параметрів регресії можна визначити за формулами:

$$t_{\beta_0} = \frac{\beta_0}{m_{\beta_0}}, \quad t_{\beta_1} = \frac{\beta_1}{m_{\beta_1}}, \quad (5)$$

або ж в Excel на рис. 1: $t_{\beta_0} = -0,130$, $t_{\beta_1} = 16,777$.

Оскільки фактичне значення критерію Стюдента для параметра β_0 менше ніж табличне значення, обране для 28 степенів свободи при рівні значущості 0,1 (вірогідності 0,9) $|t_{\beta_0}| < t_{0,1;28}$ ($0,130 < 1,7$), то параметр β_0 статистично значуще не відрізняється від 0 при цій вірогідності.

Оскільки фактичне значення критерію Стюдента для параметра β_1 більше ніж

табличне значення, обране для тієї ж кількості степенів свободи та вірогідності ($16,777 > 1,7$), то параметр β_1 статистично значуще відрізняється від 0 при цій вірогідності й гіпотезу про статистичну незначущість оцінки параметра β_1 регресії необхідно відхилити.

Довірчий інтервал оцінки параметра β_1 регресії визначається за формулою:

$$\beta_1 - t_{k;p} \cdot m_{\beta_1} \leq b_1 \leq \beta_1 + t_{k;p} \cdot m_{\beta_1}, \quad (6)$$

де b_1 – дійсне значення параметра β_1 парної лінійної регресії, яке могло б бути отримане при оцінці за генеральною сукупністю. Інструмент «Регресія» дає такий результат (рис. 1):

$$0,0189 \leq b_1 \leq 0,0232.$$

Довірчий інтервал оцінки параметра β_0 побудованої регресії визначати не варто, так як раніше нами доведено, що він статистично значуще не відрізняється від 0 при вибраній вірогідності 0,9.

І наприкінці, окрім перевірки поокремо статистичної значущості оцінок параметрів парної лінійної регресії, визначаємо якість моделі в цілому. Однією з найбільш ефективних оцінок адекватності регресійної моделі, мірою якості рівняння регресії, характеристикою її прогностичної сили є коефіцієнт детермінації, що показує частину дисперсії, пояснену рівнянням регресії.

У разі парної лінійної регресійної моделі коефіцієнт детермінації рівний квадрату коефіцієнта кореляції. Програма Excel 2010 дає нам (рис. 1) значення коефіцієнта детермінації $R^2 = 0,91$. Чим ближче R^2 до одиниці, тим краще регресія апроксимує емпіричні дані, тим ближче точки спостереження до лінії регресії. Якщо для деякої моделі $R^2 = 1$, то емпіричні точки лежать на лінії регресії і між цими змінними існує 100-відсоткова лінійна функціональна залежність, тобто без впливу жодного випадкового фактора.

Коли коефіцієнт детермінації відомий, то критерій його статистичної значущості (F -критерій Фішера) для парної лінійної регресії може бути розрахований за формулою:

$$F_{\text{факт}} = \frac{R^2 \cdot (n-2)}{1-R^2}. \quad (7)$$

або ж програма Excel 2010 дає нам його фактичне значення (рис. 1): $F_{\text{факт}} = 281,5$.

Для заданого рівня значущості α , степенів свободи k_1 та k_2 знаходимо табличне значення критерію Фішера, де $k_1 = m$ – кількість незалежних змінних у моделі, $k_2 = n - m - 1$, за допомогою статистичної функції ФРАСПОБР. Отже, критичне значення для рівня значущості $\alpha = 0,1$ та степенів свободи $k_1 = 1$, $k_2 = 28$ складає $F_{\text{кр}} = F_{0,1;1;28} = 2,89$.

Оскільки $F_{\text{факт}} > F_{\text{кр}}$ ($281,5 > 2,89$), то коефіцієнт детермінації статистично значущий при $\alpha = 0,1$.

Проведені вище розрахунки дають можливість зробити висновки про те, що побудована модель якісна та адекватна реальній дійсності (R^2 дуже близький до 1). Так як коефіцієнт детермінації дорівнює 0,91, то 91% варіації залежної змінної пояснюється отриманим рівнянням регресії, а 9% припадають на випадкові та невраховані в побудованій моделі фактори.

Останнім етапом проведення парного кореляційно-регресійного аналізу засобами MS Excel є побудова геометричної інтерпретації рівняння парної лінійної регресії, тобто лінії регресії. Взявши до уваги вихідні статистичні дані, отримані по 30-ти замірах для другої, третьої, четвертої та п'ятої контрольних точок на лінії річки, виконуємо аналогічні розрахунки засобами Microsoft Excel 2010. Результати п'яти отриманих ліній регресії зведемо в один графік, представлений на рис.4.

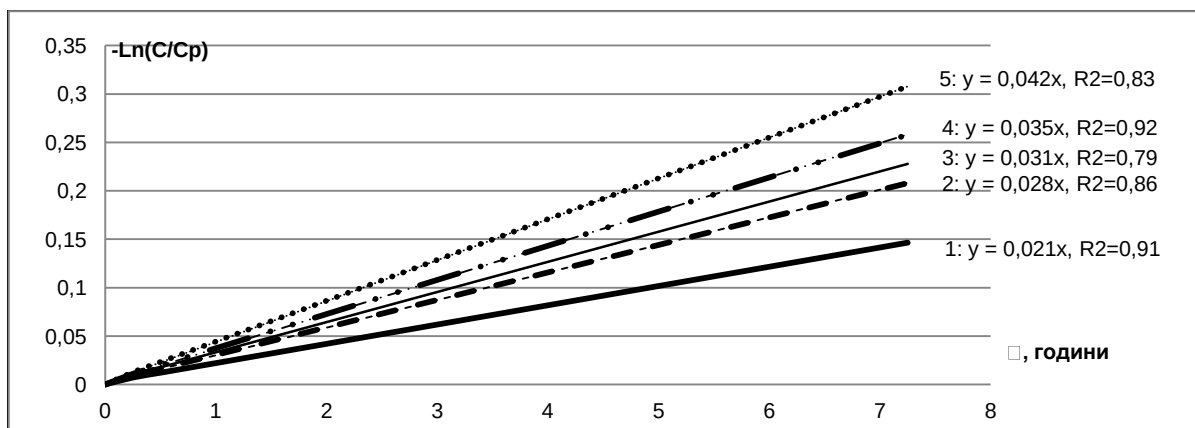


Рис. 4. Графік ліній регресії для п'яти контрольних точок

Очевидно, вони відрізняються тангенсом кута нахилу до осі часу випадіння опадів, або кутовим коефіцієнтом. Його чисельне значення визначається співвідношенням k/ε_1 , де k –

об'ємний коефіцієнт масовіддачі, ε_1 – відносний об'єм застійних зон.

Отже, для першої контрольної точки на лінії річки за отриманим рівнянням регресії (рис.

4) маємо $k/\varepsilon_1=0,021$. Звідси $\varepsilon_1=k/0,021=47,6k$. Аналогічно, для другої точки $\varepsilon_1=k/0,028=35,7k$, для третьої $\varepsilon_1=k/0,031=32,3k$, для четвертої $\varepsilon_1=k/0,035=28,6k$ та для п'ятої контрольної точки на лінії річки отримуємо $\varepsilon_1=k/0,042=23,8k$. Отже, при переході між ними по лінії річки відносний об'єм застійних зон ε_1 зменшується при фіксованому значенні об'ємного коефіцієнта масовіддачі k .

При прогнозуванні за рівнянням регресії для першої контрольної точки визначимо прогнозне значення залежної змінної y при деякому прогнозованому значенні незалежної змінної, наприклад, $x_{np}=10$ годин опадів.

Точковою оцінкою залежної змінної є її умовне математичне очікування [7], що тотожно рівне груповій середній $\hat{y}_{x_{np}}$, яку знаходимо за рівнянням регресії $\hat{y}_{np} = \beta_0 + \beta_1 \cdot x_{np} = 0,021 \cdot 10 = 0,21$.

Далі визначимо довірчий інтервал окремих можливих прогнозних значень залежної змінної $y_{x_{np}}^*$ або індивідуальних значень, які відхиляються від середньої.

Так як стандартна помилка індивідуального значення $y_{x=10}^*$ становить:

$$S_{y_{x=10}}^* = S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{np} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = 0,0149 \sqrt{1 + \frac{1}{30} + \frac{(10 - 3,625)^2}{140,4688}} = 0,0171, \quad (8)$$

то 90%-й довірчий інтервал для прогнозу індивідуального значення логарифма відносної концентрації забруднення в першій контрольній точці на десяту годину випадіння опадів буде становити:

$$\hat{y}_{x=10} - S_{y_{x=10}}^* \cdot t_{0,9;28} \leq y_{x_{np}}^* \leq \hat{y}_{x=10} + S_{y_{x=10}}^* \cdot t_{0,9;28}; \quad (9)$$

$$0,1809 \leq y_{x_{np}}^* \leq 0,2391.$$

Висновки. В статті висвітлено побудову п'яти моделей парної лінійної регресійної залежності логарифма відносної концентрації забруднення від часу випадіння опадів в різних контрольних точках заміру, доведено якість цих рівнянь регресії, що дає змогу прогнозувати хід досліджуваного процесу.

У подальших дослідження доцільно зосередити увагу на розробленні комп'ютерного програмного забезпечення для оцінювання забруднення в ґрунтовому розчині.

Список використаної літератури:

1. Євграшкіна Г. П. Гідрогеологія та основи гідромеліорації: навчальний посібник / Г. П. Євграшкіна, В. В. Войцеховська. – Дніпропетровськ: Вид-во ДНУ, 2010. – 121 с.
2. Євграшкіна Г.П., Шерстюк Н.П. Вивчення та прогнозування гідрогеологічних процесів методами математичного моделювання. Навчальний посібник / Г.П. Євграшкіна, Н.П. Шерстюк – Дніпропетровськ. Вид-во ДНУ. – 2004. – 110 с.
3. Бойко Т. В. Математичне моделювання міграції забруднюючих речовин у ґрунтах / Т.В. Бойко, А.О. Абрамова, Ю. А. Запорожець // Восточно-Европейский журнал передовых технологий. – 2013. – № 6(4). – С. 14-16.
4. Національна доповідь про стан родючості ґрунтів України. – К.: ТОВ «ВИК-ПРИНТ», 2010. – 111 с.
5. Екологічний стан геологічного і суміжних середовищ попередження аварій і катастроф на об'єктах використання надр: Звіт / Національна акціонерна компанія “Надра України”; Відповідальний виконавець Г.І. Рудько. - №12/02-1-78-02 від 12.12.2002 р. – К., 2003. – 156 с.
6. Инновационные подходы к развитию медицины, фармацевтики и эколого- биологических исследований: монография / [авт.кол.: Князева М.В., Колесов С.В., Хохленкова Н.В. и др.]. – Одеса: КУПРИЕНКО СВ, 2015 – 192 с.: ил., табл. – С.133-144, 187. – Режим доступа: <http://www.sworld.com.ua/index.php/ru/c215-8/26103-215-060>.
7. Кремер Н.Ш. Эконометрика: учебник для вузов / Н.Ш. Кремер, Б.А. Путко; под ред. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2007. – 311 с.
8. Долженков В.А. Microsoft Office Excel / В.А. Долженков, А.С. Струченков. – СПб.: БХВ-Петербург, 2011. – 816 с.
9. Лещинський О.Л. Економетрія: навч. посібн. для студ. вищ. навч. закл. / О.Л. Лещинський, В.В. Рязанцева, О.О. Юнькова. – К.: МАУП, 2000. – 208 с.

Курило А. А. Корреляционно-регрессионный анализ концентрации загрязнения в почвенном растворе инфильтрационного слоя

В статье с помощью парного корреляционно-регрессионного анализа установлена взаимосвязь концентрации загрязнения в почвенном растворе инфильтрационного слоя и времени выпадения осадков. Установлена форма зависимости между исследуемыми переменными, оценена функция регрессии и осуществлен прогноз зависимой переменной.

Ключевые слова: концентрация загрязнения, время выпадения осадков, парный корреляционно-регрессионный анализ, прогноз.

Kurylo A. O. Correlation-regression analysis of contamination concentration in the soil solution of the infiltration layer

In the article, using the pair correlation-regression analysis, the correlation of the concentration of pollution in the soil solution of the infiltration layer and the precipitation time has been established. The form of dependence between the studied variables is established, the regression function is estimated and the forecast of the dependent variable is performed.

Key words: pollution concentration, rainfall time, pair correlation-regression analysis, forecast.