Тимків Марія Михайлівна
аспірант, кафедра геотехногенної безпеки та геоінформатики
Івано-Франківський національний технічний університет нафти і газу

Тымкив Мария Михайловна
аспирант, кафедра геотехногенного безопасности и геоинформатики
Ивано-Франковский национальный технический университет нефти и газа

Tymkiv Mariia Mikhailovna
postgraduate, department of geotechnical safety and geoinformatics
Ivano-Frankivsk National Technical University of Oil and Gas

СТАТИСТИЧНА ОБРОБКА ГІДРОГЕОЛОГІЧНИХ ДАНИХ З ПРОПУСКАМИ

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ГИДРОГЕОЛОГИЧЕСКИХ ДАННЫХ С ПРОПУСКАМИ

STATISTICAL ANALYSIS OF GEOLOGICAL DATA SPACES

Анотація. Проведено статистичну обробку даних з пропусками для коректного відображення даних та точної побудови карт.

Ключові слова. Пропуски, підземні води, вихідні дані.

Аннотация. Проведено статистическую обработку данных с пропусками для корректного отображения данных и точного построения карт.

Ключевые слова. Пропуски, подземные воды, выходные данные.

Abstract. A statistical analysis of data from a space to display the correct data and accurate mapping.

Keywords. Gaps, underground water output.

Постановка проблеми. У зв'язку зі зростанням значення підземних вод у водопостачанні України, виникає необхідність постійного моніторингу підземних вод, аналізу та оцінки гідрогеологічних процесів, прогнозу можливих змін підземної гідросфери. Вивчення та прогнозування режиму рівнів підземних вод займають провідне положення в комплексі гідрогеологічних досліджень, оскільки дозволяють кількісно охарактеризувати процес формування підземних вод і прослідкувати зміни гідрогеологічних умов у часі, що дає змогу обґрунтувати заходи по використанню підземних вод, передбачити і своєчасно виявити негативні природні та техногенні впливи.

Проте, перед науковцями постає ще одна проблема: коректне опрацювання даних досліджень. Для проведення аналізу та побудови карт потрібно мати якісну вихідну інформацію. Часто складається так, що дані спостережень не є цілісними і мають багато пропусків. При певних дослідження це має значний вплив на вихідні результати та прогноз в цілому.

Одним із завдань для подальшої роботи є заповнення даних з пропусками так, щоб ці зміни коректно відображати. Коротко перелічимо деякі з основних методів для швидкого заповнення даних [3].

Методи багаторазового заповнення. Передбачають заповнення пропуску декількома значеннями. Суттєвим недоліком методів одноразового заповнення, за думкою їх дослідників є те, що звичайні формули призводять до систематично занижених оцінок дисперсії, навіть, якщо обчислені пропущені значення отримані з використанням вірної моделі. При багаторазовому заповненні отримуються правильні оцінки дисперсії, які можна отримувати звичайними методами аналізу повних даних.

Заповнення за регресією. Суть полягає в заповненні пропусків даних, які передбачені регресією пропущених значень для певного об'єкта змінних на ті, що є відомі.

Заповнення без підбору. Пропуски заповнюються постійними даними із зовнішнього джерела, наприклад даними із попередніх спостережень на цій же точці.

Заміна. Метод характеризується заміною даних на етапі спостереження та при безпосередньо при зборі даних. Суть полягає в заміні об'єкта з відсутніми потрібними даними на інший об'єкт, який не включається у вибірку.

Співставлення методів. Базується на основі декількох методів. При роботі з даними може виникнути потреба для порівняння того чи іншого методу на основі реальних спостережень. До прикладу можна об'єднати заповнення з підбором і заповнення по регресії [4].

Теоретичні відомості та математичне обґрунтування методу співставлення.

Заповнення середніми. Нехай y_{ij} - значення Y для і -го об'єкта в групі $_j$, $_i$ = 1, ..., N_j9 $_j$ = 1, ..., j. При заповненні середніми для об'єктів вибірки, які не дали відповідь, підставляється середнє y_{jR} по m_j відповіли в j-й групі. Для плану середнє популяції Y можна оцінити середнім присутніх і підставлених значень, а саме [2]:

$$\frac{\sum_{j=1}^{J} n_j \hat{y}_j}{\sum_{j=1}^{J} n_j}$$

де y_j - середнє присутніх і підставлених значень в j-й групі.

Тепер

$$\hat{y}_j = \frac{\left[m_j \bar{y}_{jR} + \left(n_j - m_j\right)\bar{y}_{jR}\right]}{n_j} = \bar{y}_{jR}$$

так що отримується оцінка Y - просто оцінка з зважуванням груп. Якщо в популяції відома частка кожної групи, то оцінку y_{ps} також можна вивести як оцінку, засновану на заповненні середніми.

Ми показали, що для планів з рівними вагами зважування об'єктів, що дають відповідь, за часткою відповідають у кожній групі дозволяє отримати такі ж оцінки середніх і сум, як підстановка середніх по відповідальним для об'єктів, що не дають відповідь. Це зауваження стосується і нерівномірним планам за умови, що вибіркові ваги відображаються в оцінках частки відповідають і в підставляється середніх. Зв'язки між заповненням пропусків і зважуванням груп розглядаються в [Oh and Scheuren A983); David, Little, Samuhel and Triest A983); Little A986)]. Метод заповнення середніми реалізується просто, але він володіє небажаними властивостями. По-перше, правильні оцінки дисперсій y_{wc} (або y_{ps}) не можна отримати за допомогою звичайних формул для дисперсії, застосованих до заповнених даними. Реально обсяг вибірки занижений через відсутність відповідей, тому звичайні формули призводять до заниженої оцінки істинної дисперсії. По-друге, величини, що не лінійні по даними, такі, як дисперсія Y або кореляція між двома змінними, не можна безбідно оцінити за допомогою стандартних методів для повних даних, якщо їх застосувати до заповнених даними. По-третє, підстановка середніх спотворює емпіричне розподіл значень Y, що важливо при дослідженні розподілу Y по гістограмі або з інших графіками, що відображає дані. Аналогічна проблема виникає, якщо значення Y об'єднані в групи для утворення частотної таблиці, тому що пропуски в групах заповнюються загальним середнім значенням і, отже, відносяться в результаті до однієї і тієї ж групи Y Ця проблема спонукає шукати розподілені значення для пропусків, використовуючи методи їх заповнення типу підстановки з підбором. Звернемося тепер до цього методу [1].

Підстановка з підбором. При більшості методів підстановки з підбором (цей термін поки не став загально прийнятим) пропуски заповнюються значеннями, отриманими для іншого східного об'єкта вибірки. Припустимо, як і раніше, що витягнута вибірка обсягу n з N об'єктів, і у m з n об'єктів вибірки зареєстровані значення Y, де n, N і m вважаються в цьому розділі фіксованими. Для простоти пронумеруємо об'єкти так, що перші n об'єктів

знаходяться у вибірws, і перші m <n з них дали відповідь. При рівноймовірній схемі вибору середнє Y можна оцінити як середнє за наявними і за підставленим значенням, що можна записати у вигляді:

$$\bar{y}_{HD} = \frac{\{m\bar{y}_R + (n-m)\bar{y}^*_{NR}\}}{n}$$

$$\bar{y}^*_{NR} = \sum_{i=1}^{m} \frac{H_i y_i}{n-m}$$

Зазначимо, що $\sum_{i=1}^{m} H_i$ рівне n-m – числу об'єктів з пропусками.

Властивості Y_{HD} залежать від способу формування чисел (H_1 ..., H_m). Найпростіше вивести формули, якщо розглядати підставлені значення як вибірку значень, отриману при ймовірному плані вибору, коли відомо розподіл (H_1 ..., H_m) при повторному застосуванні підстановки з підбором. Припустимо, що Hі задається випадковим вибором з поверненням з зареєстрованих значень Y. Умовно за зареєстрованими значенням вибірки розподіл (H_i, ..., H_m) при повторах процедури підстановки з підбором поліноміальне з об'ємом вибірки n-m і ймовірностями 1 / m, ..., 1 / m (див. [Cochran A977), розділ 2.8]. Звідси:

$$E(H_i \mid Y, R, I) = (n-m)/m,$$

$$Var\,(H_i \mid Y, R, I) = (n-m)(1-1/m)/m,$$

$$Cov\,(H_i, H_{i'} \mid Y, R, I) = -\,(n-m)/m^2,\ i \neq i'.$$

Тоді

$$E(\bar{y}_{HD1} \mid Y, R, I) = \bar{y}_R$$

$$Var(\bar{y}_{HD1} \mid Y, R, I) = \frac{(1-m^{-1})\left(1-\frac{m}{n}\right)S^2_{yR}}{n}.$$

При простому випадковому виборі і в припущенні ОПС про розподіл відповідей ми отримаємо:

$$E(\bar{y}_{HDI}\,|Y) = \bar{Y}$$

$$Var(\bar{y}_{HDI}|\,Y) = (m^{-1} - N^{-1})S_y^2 + (1 - m^{-1})\left(1 - \frac{m}{n}\right)S_y^2/n$$

Відзначимо, що підстановка з підбором веде до оцінок з більшою дисперсією в порівнянні з оцінкою Y_R одержуваної при заповненні середнього. Із формули випливає, що дисперсія будь-якої оцінки Y_{HD} при підстановці з підбором, для якої E(Y_{HD}, R, I) = Y_R більше дисперсії середнього Y_R. Перевага методу підстановки з підбором на відміну від заповнення середнім полягає в тому, що викривлення розподілу вибіркових значень відсутні. Додаткова дисперсія від вибіркової підстановки з поверненням не є малозмінена. Її можна зменшити, задаючи більш ефективний план вибору. Припустимо, наприклад, що підставляються значення витягуються без повернення. Якщо n-m <m, то ми можемо вибрати (n-m) з т зареєстрованих значень *y* без повернення і при цьому Hі дорівнює 1, якщо і -й об'єкт відібраний, і 0 - в іншому випадку. Щоб визначити процедуру в загальному випадку, запишемо:

$$n - m = km + t$$

де k - натуральне і $0 \leq t \leq m$.

При підстановці з підбором без повернення до раз вибирають всі зареєстровані об'єкти, а потім «добирають» t додаткових об'єктів, щоб забезпечити всі n - m значення, необхідних для пропусків. Таким чином,

$$\bar{y}_{NR}^{*} = \frac{km\bar{y}_R + t\bar{y}_1}{n - m},$$

де y_i- середнє t додаткових значень Y. Відповідно до теорії простого випадкового вибору

$$E(\bar{y}_1|Y,R,I) = \bar{y}_R,$$

$$Var\ (\bar{y}_1 \mid Y,R,I) = \frac{\left(1-\frac{t}{m}\right)S_{yR}^2}{t}.$$

Якщо Y_{HD2} - оцінка Y, отримана за допомогою цієї процедури, то

$$\bar{y}_{HD2} = (k+1)mn^{-1}\bar{y}_R + tn^{-1}\bar{y}_1$$

$$E(\bar{y}_{HD2}\,|Y,R,I) = \bar{y}_R$$

$$Var(\bar{y}_{HD2}|\,Y,R,I) = \frac{\left(\frac{t}{n}\right)\left(1-\frac{t}{m}\right)S_{yR}^2}{n}$$

Точніше, в припущенні простого випадкового вибору і бернулієвского розподілу присутності відповідей, ігноруючи поправку на кінцеву популяцію, ми отримаємо, що 1) дисперсія Y_{HD1} не перевищує дисперсію j_R більш ніж в 1,25 рази, і цей максимум досягається, якщо m / n - 0 , 5; 2) дисперсія j_{HD2} не перевищує дисперсії у R більш ніж в 1,125 рази. Інший метод генерування значень для заповнення пропусків - послідовний підбір, при якому всі об'єкти розташовують у послідовність і пропущене значення замінюється значенням Y найближчого попереднього в цій послідовності об'єкта, що дав відповідь. Головною перевагою послідовного підбору є його обчислювальна простота. На його основі побудовані старі схеми заповнень для поточних обстежень населення Бюро перепису (Census Bureau). Припустимо, що об'єкти вибірки випадково впорядковані і витягнуті шляхом простого випадкового вибору, а також що діє бернуліївський механізм породження пропусків. Байлар і його співавтори [див. Bailar, Bailey and Corby A978)] показали, що в цьому випадку оцінка Y

методом послідовного підбору, скажімо Y_{HD3}, - несумісна з дисперсією, наближено рівний (при великих m і n і без поправок на кінцеву популяцію)

$$Var(\bar{y}_{HD3}|\, y) = \left(\frac{S_y^2}{m}\right)\left(1 + \frac{n-m}{n}\right).$$

Значить, дисперсія Y_{HD3} збільшується в порівнянні з y_R в (n-m) / m + 1 раз, що дорівнює частці пропущених значень.

Короткий аналіз результатів. Дані, проаналізовані в даному дослідженні, отримані в результаті спостережень за водоносним четвертинним горизонтом напірного типу та з порушеним режимом поблизу м. Рівне. Проаналізовано середньомісячні значення рівня підземних вод (обчислені за даними строкових замірів) та середньомісячні значення рівня температури повітря на даній території протягом 2004-2008 років. За цими даними побудовано графіки залежності середньомісячних показників РПВ та температури відповідно за час спостережень. Обробка за допомогою методів статистичного аналізу даних моніторингу дозволила виявити деяку залежність одних показників від інших у часовому вимірі. А саме, що РПВ у вересні – жовтні, та у квітні – травні є майже однаковий, що безпосередньо пов’язано з сезонними опадами. У результаті аналізу отримали можливість побудувати чітку лінію тренду та зробити висновок, що протягом досліджуваного періоду температурний режим змінювався у циклічному порядку. Проте пряму залежність рівня підземних вод від опадів та температури стверджувати не можна. Вплив (як опосередкований, так і безпосередній) також мають сонячна активність, атмосферний тиск, рух літосферних плит, сейсмоактивність, карстові процеси та ін. Значну роль також відіграють техногенні чинники. Для прикладу наведемо графік залежності рівня підземних вод і температури повітря за жовтень протягом 2004-2008 р. р.. Обробка даних проводилась у середовищі Microsoft Excel.

Таблиця 1

Вихідні дані

Код вод. пун. за АІС ДВК Номер в.п.,що присв. партією	Тип режиму	Тип горизонту	Рік	СЕРЕДНЬОМІСЯЧНІ ЗНАЧЕННЯ РІВНЯ ПІДЗЕМНИХ ВОД (обчислені за даними строкових замірів)												Середньорічні значення рів-
				М і с я ц і												
				I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	
2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
255530005	порушений	напірний	1996			-0,74	-0,84	-0,66	-0,8	-0,76	-0,75	-0,77	-0,76	-0,74	-0,79	-0,761
			1997	-0,68	-0,73	-0,8	-0,75	-0,79	-0,78	-0,83	-0,7	-0,72	-0,72			-0,75
			1998			-0,75	-0,77	-0,74	-0,71	-0,78	-0,71	-0,73	-0,74	-0,75	-0,78	-0,746
			1999	-0,78	-0,78	-0,79	-0,72	-0,71	-0,66	-0,71	-0,71	-0,71	-0,76	-0,7	-0,71	-0,72833
			2000		-0,7	-0,69	-0,7	-0,66	-0,68	-0,72	-0,69	-0,69	-0,67	-0,7	-0,66	-0,68727
			2001	-0,66	-0,7	-0,71	-0,66	-0,67	-0,75	-0,63	-0,66	-0,66	-0,65	-0,65	-0,65	-0,67083
			2002	-0,66	-0,67	-0,67	-0,66	-0,65	-0,69	-0,68	-0,72	-0,67	-0,71	-0,7		-0,68
			2003			-0,68	-0,66	-0,68		-0,68	-0,68	-0,66	-0,73	-0,71	-0,67	-0,68333
			2004			-0,72	-0,67	-0,66	-0,69	-0,66	-0,74	-0,68	-0,5	-0,46	-0,53	-0,631
			2005	-0,46	-0,5		-0,53	-0,55	-0,51	-0,54	-0,59	-0,55	-0,54			-0,53
			2007	-1,28		-1,2	-1,18	-1,13	-1,18	-1,25	-1,18	-1,14	-1,09	-1,07	-1,07	-1,16091
			2008	-1,1	-1,09	-1,1	-1,12	-1,13		-1,17		-1,13	-1,09	-1,06	-1,13	-1,11

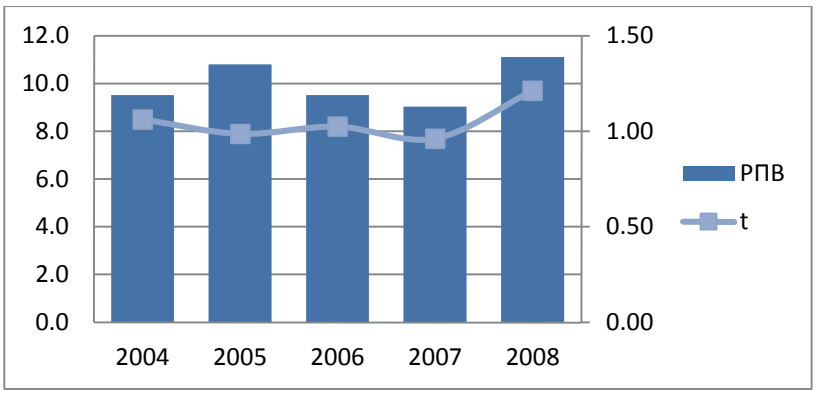

Рис. 1. Графік залежності рівня підземних вод та температури повітря у жовтні протягом 2004-2008 р.р.

Отримані коефіцієнти кореляції щодо температури повітря коливаються в межах (-0,72) - (0,6), щодо кількості опадів – (-0,77) – (0.72). Тому можемо відмітити, що статистичний зв'язок між РПВ, температурою та опадами не є однозначним протягом досліджуваного часового періоду.

Дані коефіцієнтів кореляції зміни температури повітря та кількості опадів по місяцях за 2004-2008 р. р.. наведені у таблиці 2.

Таблиця 2

Коефіцієнти кореляції зміни температури повітря та кількості опадів по місяцях за 2004-2008 р. р.

Місяці спостережень	Коефіцієнт кореляції	
	Температура повітря	Кількість опадів
Січень	0,12	-0,32
Лютий	-0,33	-0,6
Березень	-0,21	0,47
Квітень	-0,72	-0,35
Травень	0,07	-0,45
Червень	0,58	-0,55
Липень	-0,41	-0,77
Серпень	-0,30	0,72
Вересень	-0,52	0,51
Жовтень	0,62	0,35
Листопад	0,25	-0,48
Грудень	0,54	0,06

На рис. 2-3 зображені графіки залежності рівня підземних вод відносно кожного місяця. враховуючи відсутність даних за певні місяці, було проведено підстановку даних. Як видно з графіків значення R дещо збільшилось, що свідчить про доцільність даного методу. Проте він не буде коректним для великої кількості даних.

Таблиця 3

Вихідні дані до графіка зображеного на рисунку 2.

	1	2	3	4	5	6	7	8	9	10	11	12	Сер.знач.
2008	-1,1	-1,09	-1,1	-1,12	-1,13		-1,17		-1,13	-1,09	-1,06	-1,13	-1,11

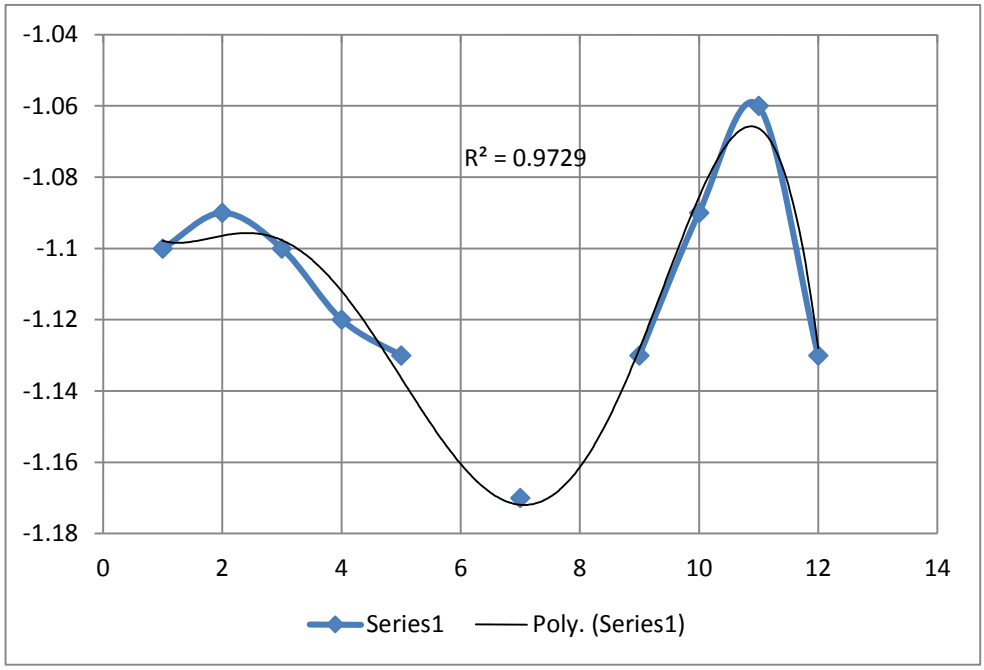

Рис. 2. Графік залежності РПВ (з пропусками)

Таблиця 4

Заповнені дані без пропусків до графіка зображеного на рисунку 3.

1	2	3	4	5	6	7	8	9	10	11	12
-1,1	-1,09	-1,1	-1,12	-1,13	-1,16	-1,17	-1,16	-1,13	-1,09	-1,06	-1,13

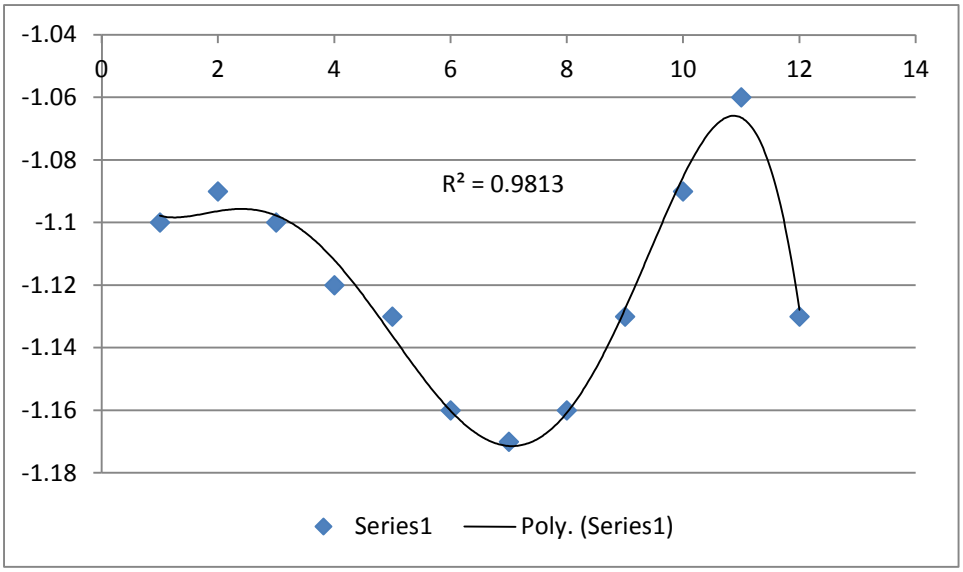

Рис. 3. Графік залежності РПВ (без пропусків)

Висновки. У процесі роботи із статистичними даними часто виникає проблема у відсутності усіх даних або у некоректному їх представлені, що не дає змогу ефективно провести аналіз та прогноз того чи іншого явища. Тому робота із пропусками на даному етапі роботи є доцільна, оскільки дослідження проводяться із щоквартальними даними.

Зокрема, в даній роботі було проведено короткий аналіз методів заповнення даних, а також представлено один із них для максимально точного відтворення даних.

Також на підставі статистичних режимних спостережень за рівнем підземних вод у межах ділянок з непорушеним режимом була встановлена певна залежність від обраних двох кліматичних чинників: середньомісячної температури повітря та середньомісячної суми атмосферних опадів. Відсутність певної залежності та залежність лиш одного або іншого фактора потребує подальших досліджень та детальнішого аналізу.

Література

1. Гандин Л.С., Каган Р.Л. Статистические методы интерпретации метеорологических данных. – Л.: Гидрометеоиздат , 1976. – 359 с.

2. Статистический анализ данных с пропусками/Пер. с англ. — М.: Финансы и статистика, 1990. — 336 с: ил.—(Математико-статистические методы за рубежом).

3. Szentimrey T. Multiple Analysis of Series for Homogenization (MASH) // Proceedings of the Second Seminar for Homogenization of Surface 53 Climatological Data, Budapest, Hungary, WMO, WCDMP-No. 41. – 1999. – P. 27-46.

4. Водообмен в гидрогеологических структурах Украины. Водообмен в естественных условиях/ [Шестопалов В. М., Дробноход Н. И., Лялько В. И. и др.]. – К.: Наукова думка, 1989. – 284 с.