УДК 551.583: 551.588

О.А. Скриник, О.Я. Скриник

ВІДНОВЛЕННЯ ПРОПУСКІВ У ЧАСОВИХ РЯДАХ МЕТЕОРОЛОГІЧНИХ ПОКАЗНИКІВ

Запропоновано метод відновлення пропусків у часових рядах метеорологічних (кліматологічних) показників. На прикладі рядів температури повітря проведено оцінку точності відновлення інформації. Проведені розрахунки показують, що точність залежить від тісноти кореляційного зв'язку між станцією, з якої відбувається відновлення, і станцією, для якої дані відновлюються. Для значень коефіцієнта кореляції $r > 0.99$ середнє квадратичне відхилення розрахованих значень від істинних не перевищує 1 °С.

Ключові слова: часові ряди, відновлення пропусків, оцінка точності.

Вступ

Для проведення кліматологічних досліджень, зокрема для побудови кліматичних карт, необхідно мати якісну вихідну інформацію, тобто, якісні ряди даних вимірювань та спостережень за основними метеорологічними величинами. Це означає, що перелік метеорологічних станцій має бути повним і репрезентативним для досліджуваної географічної території, ряди даних мають бути статистично однорідними, не містити помилок та пропусків, з максимально тривалим періодом спостережень (повинні повністю охоплювати період сучасних кліматичних змін).

Не всі з указаних умов виконуються для рядів даних, отриманих на метеостанціях України.

По-перше, існує обмежений перелік станцій з досить великим періодом спостереження (з 1881 р. і раніше), для яких дані переведено в електронний вигляд, придатний для обробки статистичного аналізу та картографування з використанням комп'ютерів.

По-друге, у зв'язку з різними причинами (переходом на іншу кількість строкових спостережень, зміною мікрокліматичних умов станцій, їх перенесенням, заміною приладів вимірювання і т.п.) на станціях порушується статистична однорідність даних. З цього приводу

зауважимо, що існують методи та прикладне програмне забезпечення, які дозволяють виявляти й усувати таку неоднорідність [5].

По-третє, проблема наявності пропусків є дуже відчутною. Практично в усіх рядах є пропуски. Їх особливо багато в кінці XIX ст., на початку XX ст., а також у період Другої світової війни, коли на багатьох метеорологічних станціях спостереження не проводились.

Мета роботи – запропонувати метод відновлення пропусків у часових рядах метеорологічних показників. Зауважимо, що всі розрахунки в роботі, зокрема й оцінка точності запропонованого методу, проводились для рядів температури повітря. Проте, як бачиться авторам, запропонований метод можна використовувати й для рядів інших метеорологічних величин з відповідним перерахунком оцінок його точності.

Формалізація задачі

Формальну постановку задачі можна представити в наступному вигляді. Розглядається сукупність часових рядів – даних синхронних спостережень за деякою метеорологічною величиною (первинних спостережень чи тих, які вже пройшли деяку попередню обробку) на певній сукупності станцій. По суті, задана матриця:

$$\left(a_{ij}\right)=\begin{pmatrix} a_{11}\ a_{12}\ a_{13}\ \ldots\ a_{1n} \\ a_{21}\ a_{22}\ a_{23}\ \ldots\ a_{2n} \\ \ldots \\ a_{m1}\ a_{m2}\ a_{m3}\ \ldots\ a_{mn} \end{pmatrix},\ i=1,2,\ldots m,\ j=1,2,\ldots n, \quad (1)$$

де m – кількість станцій, які розглядаються (кількість рядів даних спостережень), n – кількість моментів часу, в які здійснювались виміри (кількість елементів кожного ряду). Деякі елементи матриці (1) є невідомими (пропуски або лакуни).

Необхідно знайти найадекватніше значення пропущеного показника. Таке завдання можна ставити, виходячи з того факту, що рядки й стовпці матриці (1) не є статистично незалежними. Деякі з рядків (або всі вони) є залежними один від одного через те, що атмосферні процеси мають просторову протяжність (їхні лінійні масштаби відмінні від 0). Тому на достатньо близьких станціях значення однієї і тієї ж величини будуть корельованими. Деякі із стовпців теж є залежними один від одного, оскільки значення кожного окремого із рядів даних є

статистично зв’язними тією чи іншою мірою (від майже беззв’язних до зв’язних майже функціонально) через певну інерційність і циклічність атмосферних процесів.

У загальній постановці задачі можна виділити, напевне, три різних типи, які можуть зустрічатися на практиці. Перший тип – найпростіша ситуація: в деякому послідовному наборі стовпців відсутній тільки один показник. Тобто, у його локальному околі немає більше пропусків. Другий тип – в окремих рядках матриці (1) (але не у всіх) є велика кількість послідовних пропусків. Третій тип – послідовні пропуски є у всіх рядках, починаючи з деякого моменту часу. Очевидно, методи, які вирішують останній тип, можна розглядати і використовувати як методи прогнозування часових рядів.

У кліматологічній літературі прямих і ефективних методів, за допомогою яких можна відновити дані, не існує [3]. У метеорології існують методи опрацювання даних, які можна адаптувати для розв’язання поставленої задачі. Наприклад, метод просторового контролю, побудованого на оптимальній інтерполяції, який використовується для виявлення і усунення помилок спостережень. До цього ж напряму можна, напевне, віднести і методи об’єктивного аналізу [1]. Очевидно, що після певної адаптації вказані методи можна використовувати для вирішення першого типу в загальній постановці задачі відновлення даних. Для двох інших типів їх дуже важко або практично неможливо адаптувати. Зауважимо також, що існує певна кількість напрацювань щодо сформульованої задачі в кібернетиці [2]. Проте їх використання в кліматології потребує певного обґрунтування та адаптації, що не є елементарним.

Для рядів температури (середньої добової та середньої місячної), отриманих на метеорологічних станціях України, найважливішим (критичним) є другий тип із загальної постановки задачі. Крім того, враховуючи, що перший тип можна розглядати як частковий (найпростіший) варіант другого, то саме для цього типу буде здійснено опис алгоритму відновлення даних.

Алгоритм відновлення пропусків

Отже, нехай у рядку $a_{i,j}$ матриці (1) є деяка множина послідовних пропусків. На першому етапі обчислюємо коефіцієнти кореляції між

рядком $a_{i_1 j}$ і тими рядками, в яких у ці ж моменти часу пропусків немає (у деякому околі, що оточує інтервал пропусків).

Довжина (розмір) околу, в якому обчислюються кореляційні зв'язки, напевне повинна залежати від розміру множини послідовних пропусків: чим більша множина пропусків, тим більший окіл. Все ж таки, деякого граничного значення (приблизно 10-20 % від загальної кількості елементів ряду) перевищувати, напевне, не має сенсу. Окіл має складатися із множин однакового розміру (з однаковою кількістю елементів) до і після пропусків для того, щоб вловити загальну тенденцію можливих змін у рядах даних.

На другому етапі із розглядуваних рядків вибираємо той, у якого найбільший коефіцієнт кореляції із $a_{i_1 j}$ (нехай $a_{i_2 j}$). У подальшому методом найменших квадратів [4] будуємо рівняння лінійної регресії між рядками $a_{i_1 j}$ і $a_{i_2 j}$:

$$a_{i_1 j} = a \cdot a_{i_2 j} + b, \qquad (2)$$

де a і b – коефіцієнти рівняння регресії.

Для заповнення лакун необхідно лише за отриманим рівнянням і відомим значенням ряду $a_{i_2 j}$ обчислити невідомі значення ряду $a_{i_1 j}$.

Описаний алгоритм є дуже простим у практичній реалізації. Першу частину алгоритму (відшукання рядка, що найбільше корелює з даним, і побудова регресійного рівняння) можна легко і швидко виконати в пакеті прикладних статистичних програм STATISTICA. Відновлення пропущених значень за рівнянням (2) легко здійснити в табличному процесорі EXCEL.

Оцінка точності методу

Виникає питання про точність відновлення пропущених даних. Наскільки розраховані описаним методом значення відрізняються від реальних (спостережуваних) і яка кількісна оцінка цієї різниці? Очевидно, що точність відновлення даних має залежати від значення коефіцієнта кореляції r між станцією, з якої відбувається відновлення, і станцією, для якої відновлюються дані (від тісноти кореляційного зв'язку між цими станціями). Який характер такої залежності? Не тривіальним є також питання про вплив довжини відновлювального проміжку на точність відновлення.

Для вияснення сформульованих питань було проведено розрахунки для температурних рядів (середні місячні значення) за наступною

схемою. Як реперні станції, з яких проводилось відновлення, було вибрано Київ, Луганськ та Львів. Період дослідження тривалістю 20 років – з 1981 по 2000 рр. Вибір періоду був обумовлений найменшою кількістю пропусків за цей час. Для кожної з указаних станцій було сформовано перелік із 14 інших станцій, для яких здійснювалось відновлення даних. Для всього досліджуваного періоду розраховувались коефіцієнти кореляції між реперною станцією й іншими станціями та параметри регресійних рівнянь (табл. 1). Відновлення даних проводились для періодів довжиною 12, 24, 36 та 48 значень (або 1, 2, 3 та 4 роки відповідно). Відновлювані періоди вибирались таким чином, щоб до і після них була однакова кількість значень у рядах.

Таблиця 1

Оцінка точності відновлення даних у рядах середньої місячної температури станції Київ

Станція	r	$y = a \cdot x + b$		ε, $\times 10^{-3}$				σ, ^{o}C			
		a	b	12	24	36	48	12	24	36	48
Бориспіль	0,9996	1,01	-0,37	1,93	1,14	0,89	0,82	0,33	0,28	0,25	0,24
Чернігів	0,9990	1,01	-1,07	3,27	2,67	2,37	2,25	0,39	0,40	0,37	0,38
Баришівка	0,9989	1,00	-0,56	3,55	2,98	2,33	2,47	0,43	0,45	0,39	0,42
Житомир	0,9989	0,96	-0,36	2,96	2,55	2,01	1,84	0,39	0,40	0,35	0,35
Золотоноша	0,9987	1,02	-0,22	3,71	3,09	2,67	2,61	0,48	0,48	0,44	0,45
Умань	0,9983	0,98	-0,13	3,36	2,30	2,45	3,21	0,44	0,40	0,41	0,48
Конотоп	0,9981	1,04	-1,34	8,38	5,73	4,50	4,41	0,63	0,58	0,51	0,52
Вінниця	0,9981	0,96	-0,44	2,76	2,20	2,07	3,07	0,38	0,37	0,36	0,45
Шепетівка	0,9971	0,93	-0,27	5,92	4,56	3,93	4,61	0,54	0,52	0,48	0,54
Суми	0,9969	1,06	-1,74	8,62	8,33	7,05	6,78	0,62	0,68	0,62	0,63
Полтава	0,9967	1,06	-0,77	6,68	7,09	6,33	5,83	0,63	0,71	0,67	0,66
Рівне	0,9962	0,92	0,06	7,24	5,99	4,87	5,06	0,62	0,62	0,56	0,58
Дніпропетровськ	0,9937	1,07	0,01	7,65	9,96	10,6	9,93	0,75	0,93	0,93	0,94
Ів.-Франківськ	0,9912	0,90	0,48	16,1	10,1	11,0	10,8	0,96	0,83	0,86	0,88

Для кількісної оцінки точності відновлення даних було використано нормалізовану середню квадратичну помилку (ε), яка розраховується за формулою:

$$\varepsilon = \frac{\overline{(T_C - T_B)^2}}{\overline{T_C}\,\overline{T_B}}, \qquad (3)$$

де T_C – спостережуване (істинне) значення температури, T_B – відновлене значення температури, горизонтальна риска означає процедуру осереднення.

Поряд із стандартною оцінкою точності розрахунків (3) в таблиці також приведено і величину σ:

$$\sigma = \sqrt{\overline{(T_C - T_B)^2}}\,.$$

Остання дає більш наочне уявлення про міру відхилень відновлюваних значень ряду від істинних. Це, по суті, певний аналог середнього квадратичного відхилення випадкової величини.

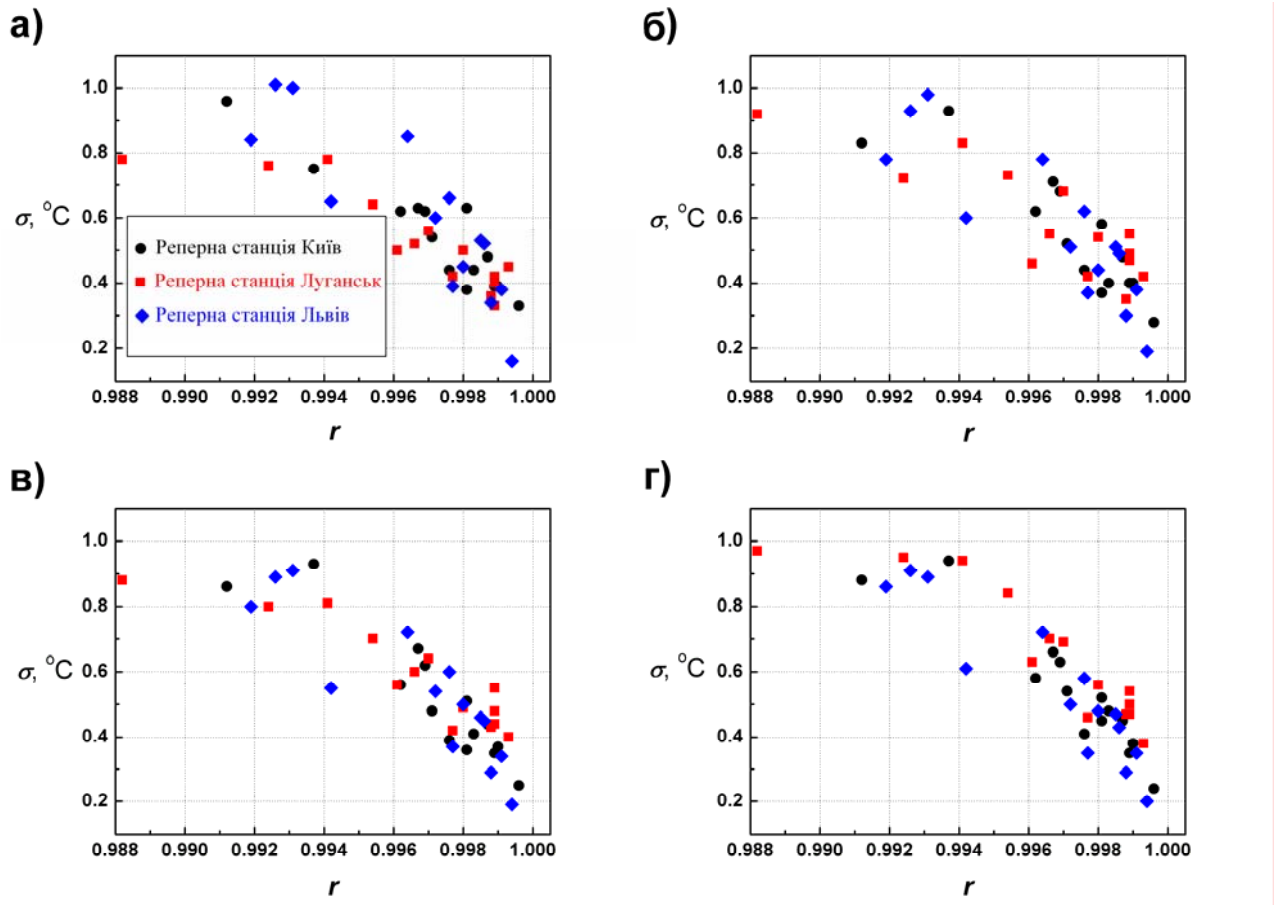

Рис. 1. Залежність оцінки точності відновлення даних від тісноти кореляційного зв’язку та довжини відновлюваного періоду: а) 12 значень; б) 24 значення; в) 36 значень; г) 48 значень

Результати розрахунків для реперної станції Київ приведено в табл. 1. Аналогічні результати отримано також і для реперних станцій Львів та Луганськ. На рис.1 представлено залежність оцінки точності σ від коефіцієнта кореляції для різних значень довжини відновлювального проміжку: 12, 24, 36 та 48 значень для трьох реперних станцій разом.

Як і слід було очікувати, зі зменшенням r точність відновлення інформації зменшується. Таке зменшення точності може бути апроксимоване простим лінійним законом. Тоді як величина проміжку пропущених значень практично не впливає на точність розрахунків.

Дані табл. 1 та аналіз графіків дозволяють зробити висновок, що для значень коефіцієнта кореляції $r > 0.99$ точність відновлення даних є досить високою, σ не перевищує 1 °C. Крім того, відхилення розрахованих значень від істинних мають практично нормальний розподіл із середнім значенням, рівним 0. Це означає, що метод не дає систематичної помилки.

Висновки

Запропонований метод є ефективним і простим у використанні. Проте важливо зауважити, що описаний алгоритм не є методом контролю якості рядів. Тобто, він буде працювати ефективно тільки в тому випадку, коли ряди, з яких здійснюється відновлення, не містять помилок (є якісними), наприклад – не містять викидів. У протилежному випадку, всі наявні помилки автоматично будуть „передані” відновлюваним даним. Тому перед застосуванням описаного алгоритму доцільно здійснити додаткову перевірку тих рядів, які планується використовувати як реперні для відновлення пропущених чи втрачених даних.

* *

1. *Гандин Л.С., Каган Р.Л.* Статистические методы интерпретации метеорологических данных. – Л.: Гидрометеоиздат, 1976. – 359 с.
2. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. – Новосибирск: изд-во ин-та математики, 1999. – 268 с.
3. *Кобышева Н.В., Наровлянский Г.Я.* Климатологическая обработка информации. – Л.: Гидрометеоиздат, 1978. – 296 с.
4. *Линник Ю.В.* Метод наименьших квадратов и основы теории обработки наблюдений. – М.: гос. изд-во физ.-мат. лит., 1962. – 350 с.
5. *Szentimrey T.* Multiple Analysis of Series for Homogenization (MASH) // Proceedings of the Second Seminar for Homogenization of Surface

Climatological Data, Budapest, Hungary, WMO, WCDMP-No. 41. – 1999. – P. 27-46.

Національний університет біоресурсів і природокористування України,
Український науко-дослідний
гідрометеорологічний інститут, Київ

О.А. Скриник, О.Я. Скриник

Восстановление пропусков во временных рядах метеорологических показателей

Предложен метод восстановления пропусков во временных рядах метеорологических (климатологических) показателей. На примере температурных рядов проведена оценка точности восстановления информации. Проведенные расчеты показывают, что точность метода зависит от тесноты корреляционной связи между станцией, с которой осуществляется восстановление, и станцией, для которой данные восстанавливаются. Для значений коэффициента корреляции $r > 0.99$ *среднеквадратическое отклонение расчетных значений от реальных (истинных) не превышает 1 °С.*

Ключевые слова: временные ряды, восстановление пропусков, оценка точности.

O.A. Skrynyk, O.Y. Skrynyk

Complementing of missing data in meteorological time series

The method which allows complementing of missing data in meteorological (climatological) time series is proposed. Using temperature time series the assessment of its accuracy was done. Calculations show that the method accuracy depends strongly on a value of a correlation coefficient between meteorological stations used for complementing. For the value of correlation coefficient $r > 0.99$ *the mean square deviation of calculated data from real data does not exceed 1 °C*

Keywords: time series, complementing of missing values, assessment of accuracy.