DOI: 10.21513/2410-8758-2019-2-14-33 УДК 551.501.6

ВЛИЯНИЕ ПРОПУСКОВ В НАБЛЮДЕНИЯХ НА ОЦЕНКУ СРЕДНЕМЕСЯЧНОЙ ТЕМПЕРАТУРЫ ВОЗДУХА

В.П. Евстигнеев [1),2)]

[1)] ФГАОУ ВО «Севастопольский государственный университет»,
Россия, 299053, г. Севастополь, ул. Университетская, 33

[2)] ФГБНУ «Институт природно-технических систем»,
Россия, 299011, г. Севастополь, ул. Ленина, 28; *vald_e@rambler.ru*

Реферат. При анализе рядов наблюдений часто возникает проблема пропусков данных. В климатологии при небольшом числе пропусков и наличии качественных рядов станций-аналогов недостающие данные восстанавливаются путем интерполяции и другими способами. В ином случае требуется специальный анализ. В настоящей работе предложен и реализован подход к оценке влияния пропусков в наблюдениях на оценку среднемесячной температуры воздуха с использованием оригинального алгоритма. Он основан на результатах проверки статистической гипотезы о равенстве выборочных средних полной и неполной выборок. Суть алгоритма сводится к тому, что в каждой из полных месячных порций среднесуточных данных генерируется набор неполных месячных совокупностей путем искусственного исключения из полной выборки заданного числа k значений подряд или вразбивку. Этот набор используется для достоверной оценки вероятности p значимого различия между среднемесячными величинами полной и неполной выборок. В случае пропусков «подряд» генерация неполных выборок выполнена путем простого смещения «окна» пропусков шириной k дней в месячной совокупности значений температуры воздуха. В случае пропусков «вразбивку» применялся один из алгоритмов неравномерной рандомизации – обобщенный алгоритм Мидзуно. Анализ вероятности p по реплицированным выборкам позволил установить пределы допустимого числа пропусков в месячных выборках и сформулировать рекомендации по оценке качества среднемесячных значений. Информационную основу работы составили ряды наблюдений за температурой воздуха на 15-ти метеорологических станциях Азово-Черноморского региона.

Ключевые слова. Метеорологические данные, пропуски, температура воздуха, контроль качества, рандомизация Мидзуно, алгоритм контроля.

Введение

Одним из ключевых требований при обработке результатов наблюдений на сети станций и постов Росгидромета является качество статистических выборок гидрометеорологических данных, позволяющих получать надежные режимные оценки, составляющие основу справочно-климатической информа-

ции для отдельных регионов. Помимо ошибок, которые практически всегда содержатся в данных, как при наблюдениях, так и при первичной обработке, на качество статистических совокупностей гидрометеорологических данных оказывает решающее влияние наличие пропусков в наблюдениях (Руководство…, 2008). В некоторых случаях пропуски можно устранять по сохраненным результатам наблюдений на твердых копиях (книжки или таблицы наблюдений) (Апарин и др., 2010) или восстанавливать либо с использованием данных близлежащих станций-аналогов (например, применяя методы приведения двух рядов к одному периоду (Кобышева, Навровлянский, 1978) и пространственной интерполяции по нескольким аналогам (Методические указания…, 1981)), либо с использованием интерполяции пропущенных значений по данным только обрабатываемой станции (Беспалов, Грошева, 1982). Однако эффективность таких способов и достоверность восстановления сильно зависит от числа пропусков в рядах и качества данных самих станций-аналогов (Massetti, 2014). В итоге обработка рядов наблюдений с пропусками для получения достоверных выводов о режиме метеорологического элемента требует специального анализа исходных данных. Предметом такого анализа при расчете среднего (за сутки, декаду, месяц) часто является возможное отличие среднего по неполному ряду от среднего по полному (Беспалов, Грошева, 1982).

В научно-методической литературе существует небольшое число работ по данной тематике, выводы которых строились на основе анализа влияния искусственно сгенерированных пропусков на оценку среднего значения в реально существующих рядах наблюдений. Так в работах (Наумова, 1983; Руководство…, 2008) моделировались неполные ряды методом Монте-Карло, и в качестве основных показателей качества использовались абсолютная или относительная погрешность определения среднемноголетней величины. Проводилось также сравнение результатов расчета при двух типах пропусков – вразбивку и подряд.

Целевой функцией в указанных исследованиях являлось среднемноголетнее значение, рассчитанное по многолетним рядам наблюдений. А вот работ, посвященных влиянию пропусков на оценку средних значений малых выборок (среднее за декаду, месяц, год), в литературе недостаточно. При этом существуют рекомендации по определению качества выборочных средних с учетом пропусков в наблюдениях, а также последовательности их возникновения в ряду (Методические указания…, 1990), однако строго объективного обоснования таких рекомендаций нет.

В связи с тем, что в последние десятилетия накоплены большие объемы метеорологических данных по отдельным регионам, автор работы посчитал целесообразным вернуться к вопросу о влиянии пропусков на расчет средних величин и определении их допустимого количества путем статистического анализа выборок данных, содержащих искусственно сгенерированные пропуски заданной длины. В настоящей работе рассмотрено влияние пропусков на оценку среднемесячного значения, рассчитанного по выборке среднесуточных величин. Основным объектом исследования стала температура воздуха в Азово-Черноморском регионе.

Постановка задачи

Проблема количества допустимых пропусков в наблюдениях сводится к вопросу о том, насколько средние значения, полученные по такой неполной выборке, близки к тем, которые могли бы быть получены по полной выборке данных. На языке статистики этот вопрос звучит так: насколько значимы различия между средними значениями, полученные по наиболее полной месячной выборке данных и по выборке, содержащей пропуски? Для проверки этой гипотезы можно воспользоваться простым параметрическим t-тестом в предположении нормальности распределения гидрометеорологических величин в месячной выборке. Конечно, этот тезис является спорным, учитывая асимметрию распределения величин в отдельные месяцы (см., например, (Костин, Покровская, 1953)), однако, по мнению авторов, такое допущение оказывается вполне разумным в первом приближении. Последующий анализ проводился без учета внутригодового распределения гидрометеорологических элементов.

Предполагается, что для каждого пункта наблюдения существует набор из N полных месячных порций данных, из которых осуществляется генерация неполных выборок путем искусственного исключения k величин. По аналогии с работой (Наумова, 1983) будем различать два рода пропусков: первого и второго рода. Пропуски вразбивку и выборки, содержащие такие пропуски, будем относить к первому роду; пропуски подряд, и соответствующие выборки – ко второму роду. Следует отметить, что на практике чаще всего приходится иметь дело с выборками второго рода. Такого типа пропуски возникают, как правило, вследствие выхода из строя прибора, в результате чего наблюдения прерываются на продолжительный период времени.

Задача определения допустимого числа пропусков сводится к вычислению частоты p_k значимых отклонений средней величины, рассчитанной по месячной выборке с k пропусками, от «истинного» среднемесячного значения. Генерируя достаточно большое число M выборок, состоящих из N «неполных» средних, можно достоверно оценить как саму вероятность p_k, так и стандартную ошибку ее расчета σ_k.

Такой алгоритм определения допустимого количества пропусков является простой реализацией метода Монте-Карло и включает в себя следующие этапы.

Этап 1. Формирование совокупности месячных выборок среднесуточных данных с отсутствием пропусков в наблюдениях.

Для каждой станции данная совокупность из N месяцев без пропусков используется, с одной стороны, в качестве базовой, расчет по которой позволяет получить N значений истинных среднемесячных величин $\tilde{M}_i$ (i=1,…,N), с другой, ─ в качестве источника генерации неполных месячных массивов, содержащих пропуски. Количество «полных» месяцев для каждой станции представлены в табл. 1. В расчетах было использовано от 666 (Сочи) до 813 (Ялта, Феодосия, Севастополь) месячных выборок, не содержащих ни одного пропуска.

Этап 2. Генерация из базовой совокупности неполных месячных порций данных.

Для каждой из N полных месячных порций данных генерируется m неполных месячных совокупностей путем искусственного исключения из полной выборки заданного числа k значений подряд или вразбивку. Для данного k рассчитывается $L=m\cdot N$ приближенных оценок средних $M_{i,j}(k)$, где $j=1,\ldots,m$.

В случае выборок второго рода для отдельно взятой месячной порции данных путем простого смещения «окна» пропусков шириной k дней может быть получен весь набор неполных выборок без необходимости использовать генераторы случайных чисел. Количество таких выборок для конкретного месяца будет равно $(D\text{-}k)$, где D – число дней в месяце.

В случае выборок первого рода для месячных порций данных число сочетаний k дней с пропусками вразбивку резко возрастает с числом k и определяется известной из комбинаторики формулой $D!/k!(D\text{-}k)!$. Уже в случае $k=3$ число таких сочетаний превышает 4000 только для одного месяца, для $k=4$ – более 27000 и т.д. На первом этапе m неполных месячных порций первого рода генерировались путем рандомизации номера суток с пропуском, используя датчик равномерно распределенных случайных чисел. Так для каждого полного месяца искусственно создавались $m=100$ неполных репликат. Простая вероятностная выборка, полученная равномерным генератором, оказалась нерепрезентативной по отношению к реальным причинам возникновения пропусков в гидрометеорологических рядах. Поэтому для достижения основной цели исследования дальнейшие вычисления проводились по результатам неравномерной рандомизации пропусков в месячных рядах. Для каждого из N «полных» месяцев равномерным датчиком генерировалось 100 номеров суток в месяце. По каждому из этих номеров центрировалась дифференциальная кривая нормального распределения со стандартным отклонением $\sigma=10$ суток и рассчитывались вероятности включения π_i. Далее из полной совокупности производился отбор значений в выборку размером k с использованием алгоритма Мидзуно (Tillé, 2006). В результате такой «двойной» рандомизации генерировались разнообразные последовательности пропусков вразбивку, неравномерно распределенных в месяце. Более подробное обсуждение метода неравномерной рандомизации представлено в Приложении.

Этап 3. Сравнение «неполных» средних с «истинными».

Для каждой из L сгенерированных месячных порций данных осуществляется проверка простой нулевой гипотезы (H_0: $M_{i,j}(k)=\tilde{M}_i$) на основе стандартного критерия $t=(M_{i,j}(k)\text{-}\tilde{M}_i)/\sqrt{V}$ в предположении, что сравниваемые выборки извлечены из одной и той же генеральной совокупности с неизвестной дисперсией. Символом V обозначена дисперсия разностей между средними значениями. Уровень значимости для двухсторонней критической области был принят равным $\alpha=10\%$, число степеней свободы $\nu = 2D\text{-}k\text{-}2$.

Присутствие в рядах среднесуточных значений температуры воздуха сериальной корреляции может существенно повлиять на результаты t-теста. Предварительные расчеты показали, что коэффициент автокорреляции в отдель-

ных случаях достигает значений 0.7-0.8. С целью учесть это влияние использован модифицированный аналог дисперсии V (Yilmaz, Aktas, 2017):

$$V_{adj} = \left| \frac{s_1^2}{2n_1}\left(1+\frac{2n_1-3}{n_1}r_1\right) + \frac{s_2^2}{2n_2}\left(1+\frac{2n_2-3}{n_2}r_2\right)\right|,$$

где $s_1{}^2$, $s_2{}^2$ – выборочные дисперсии среднесуточной температуры воздуха в «полной» и «неполной» выборках объемами n_1 и n_2, соответственно; r_1 и r_2 – коэффициенты автокорреляции первого порядка в двух сравниваемых выборках. Согласно (Yilmaz, Aktas, 2017) t-критерий с модифицированным аналогом дисперсии V_{adj} имеет большую мощность по сравнению с другими критериями в случае положительных коэффициентов r_1 и r_2.

Этап 4. Оценка частоты значимых различий между выборочными средними.

В результате применения t-теста к $M_{i,j}(k)$ можно сформировать последовательность $x^{(L)}$ длиной L, каждый элемент которой принимает значение 1 (значимое отличие в средних) с вероятностью p_k или 0 (незначимое отклонение) с вероятностью (1-p_k) в зависимости от результата t-теста. Точечной оценкой вероятности служит частота наступления события «значимое различие между средними»:

$$p_k = \frac{\sum_{i=1}^{L} x_i}{L} = \frac{r}{L} \quad . \qquad (1)$$

Таким образом, по массиву искусственно сгенерированных неполных выборок первого и второго рода производится расчет частоты значимых отклонений «неполных» средних от истинных значений по формуле (1).

Этап 5. Оценка стандартной погрешности σ_k расчета вероятности p_k.

Величина повторяемости значимых отклонений «неполных» средних от «истинных» r/L позволяет делать выводы о влиянии пропусков на оценку среднемесячных величин. Однако, для выработки рекомендаций по допустимому числу пропусков для достоверной оценки выборочной средней, необходимо учитывать выборочную ошибку расчетов, возникающую не только вследствие проблемы репрезентативности выборок, но и вследствие наличия в выборках сериальной корреляции. Вывод соответствующей формулы расчета погрешности и метод статистического оценивания подробно рассмотрены в Приложении. В конечном виде дисперсия оценки r/L определяется по формуле:

$$\mathbf{D}\left(\frac{r}{L}\right) = \frac{ab}{L(a+b)^2(a+b+1)} \cdot \left[(a+b)\left(1+\tfrac{2}{L}\sum_{j=1}^{L-1} j \frac{\Gamma(\varepsilon+\zeta)\Gamma(\varepsilon+L-j)}{\Gamma(\varepsilon)\Gamma(\varepsilon+\zeta+L-j)}\right) + L \right], \qquad (2)$$

где a, b – параметры β-распределения вероятности $p \sim Beta(a, b)$, а ε, ζ – параметры β-распределения параметра авторегрессионной модели 1-го по-

рядка $\varphi \sim Beta(\varepsilon, \zeta)$. Формула (2) позволяет оценить стандартную погрешность расчета $\sigma = \sqrt{\boldsymbol{D}\left(\frac{r}{L}\right)}$ доли «успешных» событий t-теста, а, значит, и вероятность значимого отклонения «неполного» среднего значения.

Этап 6. Определение допустимого числа пропусков первого и второго рода.

Известная величина стандартной ошибки σ_k позволяет построить односторонний доверительный интервал $p_k+\sigma_k$, который может быть использован для определения допустимого числа пропусков. Варьируя число k от 1 до 20, определяется допустимое число пропусков. В случае если $p_k+\sigma_k<=1\%$, расчет средних величин с числом пропусков k считается допустимым; $1<p_k+\sigma_k<10\%$ – получаемая по неполной выборке средняя величина сомнительна; при $p_k+\sigma_k>10\%$ – расчет средней величин недопустим при таком количестве пропусков.

Материалы исследования

Источником гидрометеорологических данных для Азово-Черноморского региона служат регулярные наблюдения, выполняемые на сети станций и постов Гидрометслужб Украины и России, обеспечивающих систематический мониторинг за состоянием атмосферы и гидросферы, сбор первичной информации, а также обработку и хранение данных по единым методикам и стандартам (Евстигнеев и др., 2014). Исследования в настоящей работе проведены по данным морских береговых станций Азово-Черноморского региона за период 1950-2013 гг., список которых приведен в табл. 1.

Таблица 1. Список станций Азово-Черноморского региона, использованных в работе

№	Станция	Код	Широта	Долгота	Высота метео-площадки (м)	Число полных месяцев, использованных в работе
1	Мариуполь	34712	47	37.5	8	767
2	Бердянск	34717	46.8	36.7	2	770
3	Очаков	33848	46.6	31.6	41	770
4	Одесса	33837	46.43	30.77	42	767
5	Геническ	33910	46.17	34.82	15	769
6	Хорлы	33917	46.1	33.3	7	760
7	Мысовое	33981	45.5	35.8	19	810
8	Опасное	33986	45.4	36.6	2	813
9	Евпатория	33929	45.2	33.4	6	810
10	Феодосия	33976	45.03	35.38	22	813
11	Геленджик	37004	44.6	38.0	27	755
12	Новороссийск	37000	44.72	37.84	30	757
13	Севастополь	33991	44.6	33.5	7	813
14	Ялта	33990	44.48	34.17	72	813
15	Сочи	37099	43.58	39.77	57	666

Информационную основу исследования составил электронный архив гидрометеорологических данных Севастопольского ЦГМС (Евстигнеев и др., 2014) по метеорологическим и морским гидрологическим параметрам, измеряемым на морской береговой сети Азово-Черноморского региона. Массив данных включает в себя характеристики гидрометеорологических элементов срочного, суточного и месячного временного масштаба.

Результаты и обсуждение

Оценка влияния числа пропусков в рядах среднесуточных значений на среднемесячную величину выполнена по данным о температуре воздуха 15-ти станций Азово-Черноморского региона с применением алгоритма, описанного в предыдущем разделе. В частности, дана точечная оценка вероятности значимых различий между средним значением, вычисленным по полной месячной выборке значений температуры воздуха, и средним значением по выборке, содержащей k пропусков подряд или вразбивку. Под значимым понималось отклонение, попадающее в критическую область α=10%. Выбор уровня значимости в целом субъективен и продиктован практикой статистических исследований. Очевидно, что при увеличении объема выборки повышается вероятность больших случайных отклонений (Яглом, 1963), поэтому выбор уровня значимости должен зависеть от объема выборки. Так, при объеме выборки, измеряемом сотнями (от 100 до 1000), уровень значимости обычно понижается до 1%, при тысячах – до 0.1%. В случае же малых объемов выборки уровень значимости может быть установлен в пределах 5-10%, что и было сделано в нашей работе.

В расчетах наиболее критичной с точки зрения достоверности результатов оказалась процедура генерации в месячных выборках искусственных пропусков вразбивку. Предварительные расчеты указали на следующее важное обстоятельство. Стандартный генератор случайных чисел распределяет пропуски по исходному месяцу равномерно. «Неполное» среднее значение оказывалось в этом случае достаточно близким к «истинному», поскольку основные особенности внутримесячного распределения среднесуточных значений температуры воздуха все же воспроизводились. В итоге, достаточно длинные, но равномерные последовательности случайных пропусков (k=15 и более) практически не сказывались на значимости различий между выборочными средними, а вероятность значимых отклонений была чрезвычайно мала. В действительности, на метеорологических станциях пропуски вразбивку возникают вследствие случайных причин, действующих не регулярно, а только в течение отдельных временных отрезков (см. рис. 1 в качестве иллюстрации). Например, при развитии неблагоприятных или штормовых погодных условий синоптического масштаба времени (3-7 суток) может на короткие сроки выходить из строя измерительный прибор с соответствующим перерывом в наблюдениях. В другой, смежный синоптический период пропуски по этим причинам, скорее всего, будут отсутствовать. Следовательно, простая вероятностная выборка, полученная равномерным генератором, ока-

зывается нерепрезентативной по отношению к реальным причинам возникновения пропусков в гидрометеорологических рядах. Это обстоятельство влечет за собой изменение условий генерации искусственных рядов и требует использование неравномерной рандомизации пропусков в месячных рядах. Идея неравномерной рандомизации заключается в формировании выборочной совокупности значений случайной переменной фиксированного размера с неравными вероятностями отбора для разных членов этой совокупности (Tillé, 2006).

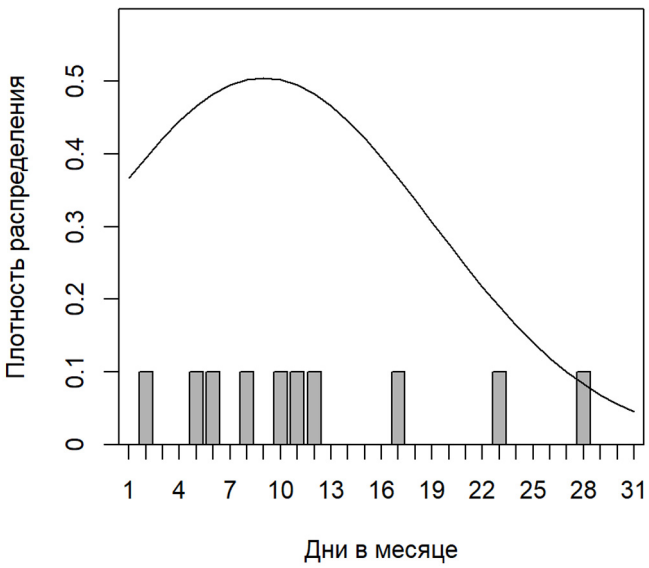
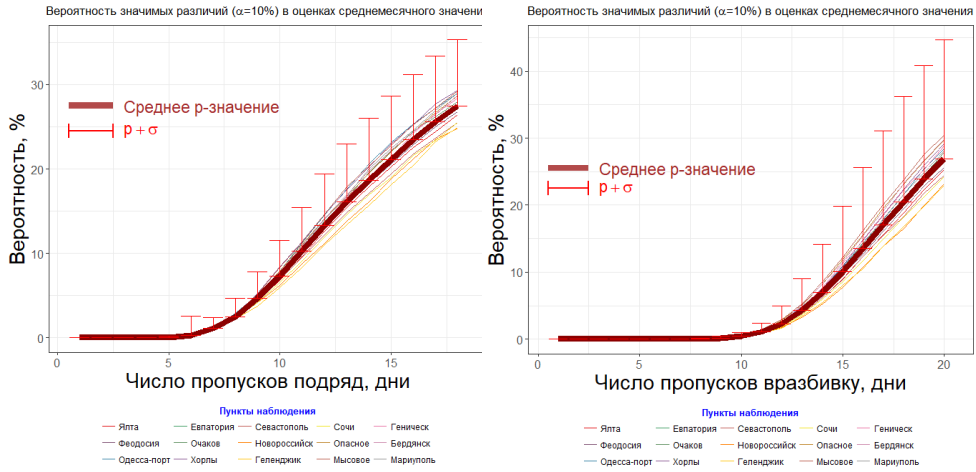


Рисунок 1. Распределение суток с пропусками вразбивку, обозначенные столбиками, и кривая плотности нормального распределения $N(m=9, \sigma=10)$

Результат расчета вероятности значимых отклонений между средним значением, вычисленным по полной месячной выборке значений температуры воздуха, и средним значением по выборке, содержащей k пропусков подряд или вразбивку, представлен на рис. 2 в виде кривых p_k. Кривые монотонны, достаточно гладки и имеют сигмоидную форму. Гладкость кривых определяется тем, что для расчетов использовалось большое количество искусственно реплицированных выборок с пропусками. Так было сгенерировано от 45000 до 81000 «неполных» выборок, содержащих пропуски вразбивку, от 6000 до 37500 выборок с пропусками подряд в зависимости от количества «полных» месяцев для отдельных станций и длины последовательности пропусков.

Из рис. 2 видно, что вероятность значимых отклонений при заданном k подчас в разы выше для пропусков подряд (второго рода), чем для пропусков вразбивку (первого рода). Так при k=15 суток, вероятность отклонений p_{15} составила 18.1-23.2% для пропусков второго рода, что почти в 2 раза выше чем для пропусков первого рода 7.7-12.3% при том же k. Наличие последовательных пропусков действительно критично, поскольку они «маскируют» целые периоды изменения температуры, обусловленные развитием синопти-

ческих процессов, в том числе приводящим к появлению максимальных или минимальных ее значений (например, атмосферные волны тепла/холода). Качественно отмеченная закономерность совпадает с результатами других работ (Наумова, 1983; Руководство…, 2008; Massetti, 2014), в которых также отмечается высокая чувствительность средних характеристик к присутствию пропусков подряд.



а) б)

Рисунок 2. Вероятность p_k значимых (α=10%) отклонений в оценках среднемесячной температуры воздуха в зависимости от числа пропусков k подряд (а) или вразбивку (б)
Стандартная погрешность σ расчета получена с учетом сериальной коррелированности в выборке результатов t-теста

Во взаимном распределении кривых p_k для разных метеорологических станций замечена некоторая закономерность. Чаще всего, наибольшую p_k имели станции Хорлы, Мысовое, Одесса, Бердянск, Мариуполь для выборок первого и второго рода, наименьшая p_k почти всегда отмечалась для станций Новороссийск, Геленджик, Сочи, Ялта, Севастополь. В работах (Лемешко и др., 2014; Евстигнеев и др., 2016) были отмечены региональные особенности распределения линейных трендов температуры воздуха Азово-Черноморского региона и показателя ее внутримесячной изменчивости (межквартильное расстояние). Причем группа станций, располагающаяся в зоне субтропического климата (от Сочи до Севастополя), имела распределение трендов, существенно отличавшееся от группы станций в умеренной климатической зоне. По всей видимости, аналогичная «региональность» проявляется во взаимном расположении кривых p_k.

Решающее влияние на результат статистического t-теста, а, значит, и на p_k, оказывает выборочное стандартное отклонение s сравниваемых месячных выборок. Можно предположить, что между величинами p_k и s существует формальная статистическая связь. Для грубой оценки этой связи были получены среднемноголетние величины $\bar{s}$ для каждой из 15 станций. По совокупности 15-ти значений $\bar{s}$ и p_k для заданного числа пропусков k рассчитан коэффициент

корреляции r_{sp}. В результате получено, что при k>9 пропусках вразбивку, r_{sp} достигло значений 0.5-0.6; при k>9 пропусках подряд r_{sp} достигало значений 0.6-0.7. Другими словами, вероятность отклонений p_k увеличивается с ростом s, и наоборот, для метеорологических станций, где внутримесячная изменчивость незначительна, влияние пропусков на оценку среднемесячной температуры воздуха снижается. Роль «климатического» месячного стандартного отклонения также была отмечена в работе (Massetti, 2014), одним из выводов которой стала рекомендация использовать ее в качестве дополнительного источника информации о степени потенциального влияния пропусков на надежность климатологической обработки данных и хранить ее величину наряду с другими метаданными метеорологических станций.

Несмотря на отмеченную зависимость, в целом кривые p_k для 15-ти станций образуют достаточно плотную группировку, что дает основание считать результаты расчетов для разных станций выборкой из одной генеральной совокупности, общей характеристикой которой может быть средневзвешенное математическое ожидание и дисперсия:

$$\overline{p}_k = \frac{\sum_{i=5}^{15} n_i p_k^{(i)}}{\sum_{i=5}^{15} n_i} \qquad \overline{\sigma}_k^{\,2} = \frac{\sum_{i=5}^{15} n_i \sigma_k^{(i)2}}{\sum_{i=5}^{15} n_i} \quad .$$

На рис. 2 кривые средневзвешенных вероятностей нанесены утолщенной линией. Средневзвешенная вероятность может быть использована для выработки рекомендаций по допустимому числу пропусков в месячных выборках и простановке признаков качества. Однако более верным будет использование границ интервала $\overline{p}_k + \overline{\sigma}_k$, которые также указаны на рис. 2. Безусловно, выбор предельных значений p_k+σ_k, соответствующих признакам качества, субъективен, однако в нашей работе он был подчинен некоторой логике. Например, допустимое число пропусков в ряду должно определяться настолько малой вероятностью p, чтобы значимое отклонение «неполного» среднего от своего «истинного» значения было маловероятным событием. С учетом того, что реальное число месяцев с пропусками в существующих климатических рядах данных обычно мало, допустимая граница вероятности p в 1% показалась нам вполне разумной. Предельное число пропусков k, определяющих «неполное» среднее $M_{i,j}(k)$ как «сомнительное», имеет более широкий интервал значений p. Анализ результатов вычислений по данным отдельных станций показал, что для p=1-5% количество пропусков k, соответствующих таким p, отличается в пределах 1-2 суток. Для увеличения надежности индикации «сомнительных» средних, предел p был повышен до 10%.

На основе этих границ сформулированы основные рекомендации для пропусков первого и второго рода, представленные в табл. 2. Так, допустимым числом пропусков подряд, практически не влияющим на оценку выборочного среднемесячного значения, можно считать от 1 до 5 суток; для пропусков вразбивку – от 1 до 9 суток. В остальных случаях среднемесячные значения должны быть забракованы, либо приняты с признаком «сомнительно».

Таблица 2. Допустимое число пропусков в рядах температуры воздуха в месячной выборке и рекомендуемые признаки качества

Кол-во пропусков	Признак качества		
	Допустимо $p \leq 1\%$	Сомнительно $1\% < p < 10\%$	Недопустимо $p \geq 10\%$
Подряд	5 и менее	6-9	10 и более
Вразбивку	9 и менее	10-13	14 и более

Заключение

Полученные рекомендации справедливы для Азово-Черноморского региона. Распространение рекомендаций на другие регионы должно осуществляться с осторожностью, поскольку, как было отмечено выше, величина p_k находится в некоторой зависимости от характеристики изменчивости температуры воздуха (стандартного отклонения) в регионе. В случае если стандартное отклонение температуры воздуха принимает низкие значения, число допустимых пропусков в рядах метеорологических станций такого региона может оказаться заметно больше указанного в табл. 2.

Здесь стоит отметить, что по данным работ (Лемешко и др., 2014; Евстигнеев и др., 2016) внутримесячная изменчивость температуры воздуха имеет сезонную специфичность. В дальнейшем, вероятно, потребуется выполнить анализ вероятностей p_k значимых различий в выборочных оценках среднего при наличии пропусков для каждого календарного сезона (месяца) по отдельности.

В заключении важно подчеркнуть, что проблема влияния пропусков в рядах наблюдений сохраняет свою актуальность и признается в том числе мировым научным сообществом. В частности, в (Massetti, 2014) проведено серьезное исследование влияния пропусков подряд и вразбивку на оценку среднемесячных значений температуры воздуха на большом фактическом материале (более 3000 станций по всему миру). Предложенная в работе модель оценки максимальной ошибки выборочного среднего в зависимости от числа дней с пропусками и стандартного отклонения температуры воздуха в месяце определена в абсолютных величинах (°C). Для станций с высокой внутримесячной изменчивостью влияние ошибки в 0.1-0.2°C будет несущественным и месяцы, содержащие пропуски в пределах такой ошибки, могут быть использованы в климатологическом исследовании. Для станций с незначительной изменчивостью температуры воздуха влияние ошибки в 0.1-0.2°C может быть критичным.

В такой ситуации возрастает неопределенность для выработки общих рекомендаций по оценке качестве расчетных данных, полученных на основе «неполных» массивов.

В данной же работе была сформулирована и решена задача влияния пропусков с использованием более общего подхода, основанного на результатах проверки статистической гипотезы о равенстве выборочных средних. Предло-

женные на его основе рекомендации имеют более обобщающий характер и могут быть использованы на практике.

Благодарность

Автор выражает благодарность Лемешко Наталье Александровне за полезные советы и критические замечания по форме и содержанию представленного в работе материала.

Работа выполнена при поддержке фонда РФФИ в рамках проекта № 18-05-01073.

Приложение

Метод генерации неполных месячных порций данных с использованием неравномерной рандомизации

Задача формирования выборочной совокупности значений случайной переменной путем неравномерной рандомизации состоит в следующем. Пусть имеется общая совокупность значений размером n (в условиях поставленной задачи n=28,29…31). Из этой общей совокупности извлекается выборка размером $k<n$. Причем i-й член общей совокупности имеет разную, отличную от 0, вероятность π_i (i=1…n) включения в выборку k. Рандомизация может производиться либо с повторением (выборка размера k формируется k отборами из полной совокупности, в результате чего отдельные его члены могут быть отобраны несколько раз), либо без повторения. Исходя из поставленной задачи, следует использовать вариант рандомизации без повторений, в результате чего формируется выборка размером k, содержащая k уникальных единиц полной совокупности, т.е. k уникальных номеров суток в месяце.

Как правило, отличие в вероятностях включения π_i связана с разным «размером» или весом i-й единицы полной совокупности. Размер можно выразить некоторым положительным числом, пропорциональным вероятности π_i. Поскольку отсутствует какая-либо другая априорная информация, в качестве такого положительного числа в настоящей работе использовалась плотность нормального распределения с параметрами: m = некоторый номер суток в месяце, σ=10 суток. Соответственно, сутки с номерами близкими к m имеют больше шансов попасть в выборку, тем самым, моделируется нерегулярность возникновения пропусков вразбивку. Параметр σ был грубо оценен, исходя из анализа стандартного отклонения дат (номера дня в месяце) реальных пропусков в рядах наблюдений относительно центральной даты (номера дня в месяце) всех пропусков в каждом месяце. По данным 15-ти станций параметр σ оказался равным 7-8 суток. Учитывая малое число месяцев с пропусками, было принято решение увеличить σ до 10 суток, хотя конечный результат для σ от 7 до 10 суток варьировался незначительно.

Существует множество алгоритмов неравномерного отбора, отличающихся стратегией формирования выборки с учетом заданных вероятностей включения π_i. Подробный обзор существующих алгоритмов неравномерной

рандомизации представлены в книге (Tillé, 2006). В настоящих расчетах использовался один их простых методов рандомизации – обобщенный алгоритм Мидзуно. Как показали предварительные расчеты, другие методы, такие как, например, Тиле и метод систематического отбора, дают примерный схожий результат.

Для каждого «полного» месяца равномерным датчиком случайных чисел генерировалось 100 номеров суток в месяце. По каждому из этих номеров центрировалась дифференциальная кривая нормального распределения со стандартным отклонением σ=10 суток и рассчитывались вероятности включения π_i. Далее из полной совокупности производился отбор значений в выборку размером k с использованием алгоритма Мидзуно (Tillé, 2006). В результате такой «двойной» рандомизации генерировались разнообразные последовательности пропусков вразбивку, неравномерно распределенных в месяце.

Формула расчета стандартной погрешности σ_k и способ статистического оценивания

В результате применения t-теста к L парам «полных» и «неполных» выборок можно сформировать последовательность $x^{(L)}$, каждый член x_i которой является бинарной случайной величиной, принимающей значения 1 (значимое отклонение в средних – «успех») с вероятностью p_k или 0 (незначимое отклонение – «неудача») с вероятностью (1-p_k). Число «успехов» $r = \sum_{i=1}^{L} x_i$ в последовательности $x^{(L)}$ подчиняется биномиальному закону распределения.

$$Bin\left(r \middle| p_k, L\right) = \binom{L}{r} p_k^r \left(1 - p_k\right)^{L-r} \qquad (\text{П.1})$$

Таким образом, погрешность статистической оценки p_k определяется свойствами биномиальной случайной величины. В дальнейшем индекс k будет опущен для простоты представления формул.

Формула (П.1) справедлива в случае, если вероятность p является неслучайной величиной и достоверно известна. Однако, требовать этого, в действительности, нельзя, – строго говоря, сама p имеет некоторое распределение. Для величин $p \in (0,1)$ часто (MacKay, 2003) применяется двухпараметрическое β-распределение $p \sim Beta\ (a,b)$, которое регулируется параметрами a, b. Несмотря на существующую критику аргументированности такого подхода (см. п. 5 гл. 25 в (Джонсон и др., 2012)), в настоящей работе также использовалось β-распределение, поскольку какая-либо дополнительная информация о виде распределения p отсутствовала.

Если предположить наличие сериальной коррелированности в ряду результатов t-теста, то в первом приближении связанность членов последовательности может аппроксимироваться простой авторегрессионной моделью 1-го порядка, параметр φ которой численно равен коэффициенту корреляции между смежными членами ряда и определен на отрезке $\varphi \in (0,1)$. Аналогично

p, выборочная оценка параметра φ должна иметь некоторое априорное распределение. Для аппроксимации распределения φ используют приближенную формулу Лейпника, которая, тем не менее, является некоторым вариантом семейства симметричных β-распределений (Джонсон и др., 2012). По этой причине, а также для унификации изложения материала и упрощения математических выкладок, в настоящей работе также использовалось β-распределение $\varphi \sim Beta\ (\varepsilon, \zeta)$ для коэффициента авторегрессии.

Предполагая, что выборочные параметры могут быть найдены из данных неравномерной рандомизации, рассмотрим формулу для расчета дисперсии оценки r/L.

Если r/L, p, φ являются случайными величинами и определены в одном и том же вероятностном пространстве, то полная дисперсия (r/L) может быть рассчитана по формуле (Bowsher, Swain, 2012):

$$\mathbf{D}\left(\frac{r}{L}\right) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L}\middle| p,\varphi\right)\right] + \mathbf{E}\left[\mathbf{D}\left(\mathbf{E}\left[\frac{r}{L}\middle| p,\varphi\right]\middle| p\right)\right] + \mathbf{D}\left[\mathbf{E}\left(\frac{r}{L}\middle| p\right)\right],$$

откуда немедленно следует:

$$\mathbf{D}\left(\frac{r}{L}\right) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L}\middle| p,\varphi\right)\right] + \mathbf{E}\left[\mathbf{D}\left(p\middle| p\right)\right] + \mathbf{D}\left(p\right) = \mathbf{E}\left[\mathbf{D}\left(\frac{r}{L}\middle| p,\varphi\right)\right] + \mathbf{D}\left(p\right) \quad . \text{(П.2)}$$

В первое слагаемое в правой части ур. (П.2) входит дисперсия выборочного среднего при наличии сериальной корреляции. При большом объеме выборки она рассчитывается по формуле (Wilks, 1997):

$$\mathbf{D}\left(\frac{r}{L}\middle| p,\varphi\right) = V \cdot \frac{s_x^2}{L} = \frac{1}{L} \cdot p(1-p) \cdot V \quad .$$

Здесь множитель V учитывает соответствующее смещение выборочной дисперсии среднего и связан с коэффициентом авторегрессии 1-го порядка φ:

$$V = 1 + 2\sum_{j=1}^{L-1} \frac{j}{L} \varphi^{L-j} \quad .$$

Продолжив преобразования (П.2) с учетом β-распределений p, φ, получим:

$$\mathbf{D}\left(\frac{r}{L}\right) = \mathbf{E}\left[\frac{1}{L} \cdot p(1-p) \cdot V\right] + \mathbf{D}\left(p\right) = \frac{1}{L} \cdot \mathbf{E}\left[p - p^2\right] \cdot \mathbf{E}\left(V\right) + \mathbf{D}\left(p\right) =$$

$$= \frac{1}{L} \cdot \left[\mathbf{E}\left(p\right) - \mathbf{D}\left(p\right) - \mathbf{E}^2\left(p\right)\right] \cdot \left(1 + \tfrac{2}{L}\sum_{j=1}^{L-1} j \frac{\Gamma(\varepsilon+\zeta)\Gamma(\varepsilon+L-j)}{\Gamma(\varepsilon)\Gamma(\varepsilon+\zeta+L-j)}\right) + \mathbf{D}\left(p\right)$$

Поскольку формулы для расчета математического ожидания и дисперсии величин, подчиняющихся β-распределению, известны (Джонсон и др., 2012), то после очевидных упрощений приведем окончательный вид расчетной формулы для дисперсии r/L:

$$\mathbf{D}\left(\frac{r}{L}\right)=\frac{ab}{L(a+b)^2(a+b+1)}\cdot\left[(a+b)\left(1+\tfrac{2}{L}\sum_{j=1}^{L-1}j\frac{\Gamma(\varepsilon+\zeta)\Gamma(\varepsilon+L-j)}{\Gamma(\varepsilon)\Gamma(\varepsilon+\zeta+L-j)}\right)+L\right]. \quad (\text{П.3})$$

Формула (П.3) позволяет оценить стандартную погрешность расчета $\sigma = \sqrt{\boldsymbol{D}\left(\frac{r}{L}\right)}$ доли «успешных» событий t-теста, а, значит, и вероятности значимого отклонения «неполного» среднего значения.

Точечная оценка параметров распределений, используемых для расчета стандартной погрешности σ при k пропусках, может быть дана с помощью метода максимального правдоподобия применительно к сумме результатов испытаний t-теста r. Для этого, вообще говоря, необходимо построить функцию правдоподобия на пространстве некоторой совокупности выборочных значений $\tilde{r}$. С этой целью $x^{(L)}$ был поделен на N последовательностей $x^{(n)}{}_l$ (l=1..N) меньшей длины $n<L$ (в расчетах n было принято равным 200 для обоих типов пропусков, а N=[L/n]).

Если принять во внимание законы распределения величин r и p, то безусловным распределением r будет:

$$f(r)=\int_0^1 f(r|p)\cdot f(p)dp=\int_0^1 Bin(r|p,n)\cdot Beta(r|a,b)dp$$

т.е. бета-биномиальное распределение:

$$f(r)=BB(r|n,a,b)=\frac{\Gamma(a+b)\Gamma(a+r)\Gamma(n-r+b)\Gamma(n+1)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)\Gamma(r+1)\Gamma(n-r+1)} \quad (\text{П.4})$$

На основе функции (П.4) строится искомая функция правдоподобия,

$$LK(a,b)=\prod_{l=1}^{N}BB(r_l|n,a,b),$$

максимизация которой позволяет получить оценки параметров

$$(\hat{a},\hat{b})=\arg\max_{a,b} LK\ (a,\ b).$$

По набору последовательностей $x^{(n)}{}_l$ может быть дана точечная оценка параметров $(\hat{\varepsilon},\hat{\zeta})$ β-распределения коэффициента авторегрессии φ тем же методом максимального правдоподобия. Для этого достаточно построить функционал:

$$LK(\varepsilon,\zeta)=\prod_{l=1}^{N}Beta(\varphi_l|\varepsilon,\zeta),$$

где φ_l – коэффициент сериальной корреляции при единичном лаге, оцененный по последовательности $x^{(n)}{}_l$.

Поиск максимума функций правдоподобия и определение необходимых параметров может быть выполнен градиентным методом, например, методом Ньютона-Рафсона. В заключении стоит отметить, что в расчетах предпочтительней пользоваться логарифмом функции правдоподобия $\ln LK$.

Список литературы

Апарин Б.В., Веселов В.М., Сивачок С.Г. 2010. Устранение пропусков в рядах метеорологических наблюдений. – Труды ВНИИГМИ-МЦД, вып. 174, с. 8-13.

Беспалов Д.Я., Грошева Л.А. 1982. О восполнении пропущенных данных в материалах наблюдений. – Труды ГГО, вып. 461, с. 19-24.

Джонсон Н.Л., Коц С., Балакришнан Н. 2012. Одномерные непрерывные распределения: Ч. 2 / пер. 2-го англ. изд. – М., БИНОМ, 600 с.

Евстигнеев В.П., Наумова В.А., Евстигнеев М.П. Лемешко Н.А. 2016. Физико-географические факторы сезонного распределения линейных трендов температуры воздуха на примере Азово-Черноморского побережья. – Метеорология и гидрология, № 1, с. 29-40.

Евстигнеев В.П., Евстигнеев М.П., Кульбида Н.И., Наумова В.А., Швень Н.И., Мосунова Е.А. 2014. Создание унифицированной базы метеорологических данных Украины. – Труды ВНИИГМИ-МЦД, вып. 178, с. 175-184.

Кобышева Н.В., Навровлянский Г.Я. 1978. Климатологическая обработка метеорологической информации. – Л., Гидрометеоиздат, 295 с.

Костин С.И., Покровская Т.В. 1953. Климатология. – Л., Гидрометеоиздат, с. 269-276.

Лемешко Н.А., Евстигнеев В.П., Наумова В.А. 2014. Изменения температуры воздуха в Азово-Черноморском бассейне и на территории Крыма. – Вестник СПбГУ, Серия 7, вып. 4, с. 131-143.

Методические указания по проведению критического контроля результатов метеорологических наблюдений на сети станций. 1981. – Л., Гидрометеоиздат, 71 с.

Методические указания. Обработка и контроль данных прибрежных гидрологических наблюдений морских береговых гидрометеорологических станций и постов. – РД 52.10.216-89. 1990. – М., Гидрометеоиздат, 140 с.

Наумова Л.П. 1983. Оценка влияния пропусков наблюдений на значения климатических характеристик. – Труды ГГО, вып. 475, с. 20-25.

Руководство по специализированному климатическому обслуживанию экономики. 2008. /Под ред. Н.В. Кобышевой. – СПб., 336 с.

Яглом А.М. 1963. Статистические методы экстраполяции метеорологических полей – В кн.: Тр. Всесоюз. научн. метеорол. совещ. «Динамическая метеорология». – Л., Гидрометеоиздат, 280 с.

Bowsher C.G., Swain P.S. 2012. Identifying sources of variation and the flow of information in biochemical networks. – Proc. Natl. Acad. Sci. USA, vol. 109(20), pp. 1320-1329.

MacKay D.-J.C. 2003. Information Theory, Inference, and Learning Algorithms. – Cambridge University Press, 628 p.

Massetti L. 2014. Analysis and estimation of the effects of missing values on the calculation of monthly temperature indices. – Theor. Appl. Climatol., vol. 117, iss. 3-4, pp. 511-519.

Tillé Y. 2006. Sampling Algorithms. – Springer Series in Statistics. New York, Springer, 216 p., doi: 10.1007/0-387-34240-0.

Wilks D. 1997. Resampling hypothesis tests for autocorrelated fields. – J. Clim., vol. 10, pp. 65-82.

Yilmaz A.E., Aktas S. 2017. Autocorrelation Corrected Standard Error for Two Sample t-test Under Serial Dependence. – Hacettepe Journal of Mathematics and Statistics, vol. 46(6), pp. 1199-1210, doi: 10.15672/HJMS.201611515847.

Статья поступила в редакцию: 21.12.2018 г. *После переработки: 12.03.2019 г.*

INFLUENCE OF MISSING DATA ON THE ESTIMATION OF MONTHLY MEAN AIR TEMPERATURE

V.P. Evstigneev [1),2)]

[1)] Sevastopol State University,
33, Universitetskaya str., 299053, Sevastopol, Russian Federation

[2)] Institute of Natural and Technical Systems,
28, Lenin str., 299011, Sevastopol, Russiaт Federation; *vald_e@rambler.ru*

Abstract. While making analysis of standard observational meteorological data a problem of missing data arises. In climatology missing values are recovered by interpolation or another method if the number of missing values is small and high-quality dataset from representative reference station is available. Otherwise, special treatment is needed for initial data. In the present study an influence of missing values on monthly mean air temperature has been examined using original algorithm. It is based on the results of a two-sample t-test of the difference between means of the incomplete and complete populations. The essence of the algorithm is that a set of incomplete populations is produced for each complete monthly population of daily values by means of consecutive or random elimination of *k* values. The set is used for valid estimation of probability *p* of significant difference between complete and incomplete monthly means. In case of consecutive missing values, generation of incomplete data was accomplished by simple moving of *k*-gaps 'window' along daily air temperature data for the month. In case of random gaps, one of the unequal probability randomization algorithms was applied – generalized Midzuno method. Analysis of the probability *p* estimated over replicated populations enabled to determine the limits of permitted number of gaps and to give recommendations for quality control of mean monthly values. The study was performed on air temperature time series observed at 15 meteorological stations in the Azov and the Black Sea region.

Keywords. Meteorological data, missing values, air temperature, quality control, Midzuno randomization, quality control algorithm.

References

Aparin B.V., Veselov V.M., Sivachok S.G. 2010. Ustranenie propuskov v ryadah meteorologicheskih nablyudenij [Reconstruction of missing values in timeseries of meteorological observations]. *Trudy VNIIGMI-MCD – Proc. All-Russian Resh Institute of Hydrometeorological Information – World Data Center*, vol. 174, pp. 8-13.

Bespalov D.Ya., Grosheva L.A. 1982. O vospolnenii propushchennyh dannyh v materialah nablyudenij [On gap-filling of the missing data in observations]. *Trudy GGO – Proc. Voeikov Main Geophysical Observatory*, vol. 461, pp. 19-24.

Dzhonson N.L., Koc S., Balakrishnan N. 2012. *Odnomernye nepreryvnye raspredeleniya: Ch.2 / per. 2-go angl. izd.* [Continuous Univariate Distributions: vol.2 / trans. 2nd ed.]. Moscow, Binom, 600 p.

Evstigneev V.P., Naumova V.A., Evstigneev M.P. Lemeshko N.A. 2016. Physiographic factors of seasonal distribution of linear trends in air temperature on the Azov-Black sea coast. [Physical and geographical factors of the seasonal distribution of linear air temperature trends on the example of the Azov-Black Sea coast]. *Meteorologiya i gidrologiya – Meteorology and Hydrology*, no. 1, pp. 29-40.

Evstigneev V.P., Evstigneev M.P., Kul'bida N.I., Naumova V.A., Shven' N.I., Mosunova E.A. 2014. Sozdanie unificirovannoj bazy meteorologicheskih dannyh Ukrainy [Creation of a unified database of meteorological data of Ukraine]. *Trudy VNIIGMI-MCD – Proc. All-Russian Resh Institute of Hydrometeorological Information – World Data Center*, vol. 178, pp. 175-184.

Kobysheva N.V., Navrovlyanskij G.Ya. 1978. *Klimatologicheskaya obrabotka meteorologicheskoj informacii* [Climatological processing of meteorological data]. Leningrad, Gidrometeoizdat, 295 p.

Kostin S.I., Pokrovskaya T.V. 1953. *Klimatologiya* [Climatology]. Leningrad, Gidrometeoizdat, pp. 269-276.

Lemeshko N.A., Evstigneev V.P., Naumova V.A. 2014. Izmeneniya temperatury vozduha v Azovo-Chernomorskom bassejne i na territorii Kryma [Air temperature changes on the Azov-Black sea coast and the Crimea peninsula]. *Vestnik Sankt-Peterburgskogo Universiteta, Seriya Geologiya i Geografiya – Bulletin of Saint-Petersburg State University, Geology and Geography*, vol. 4, pp. 131-143.

Metodicheskie ukazaniya po provedeniyu kriticheskogo kontrolya rezul'tatov meteorologicheskih nablyudenij na seti stancij [Instructions for quality control of observational data on meteorological stations]. 1981. Leningrad, Gidrometeoizdat, 71 p.

Metodicheskie ukazaniya. Obrabotka i kontrol' dannyh pribrezhnyh gidrologicheskih nablyudenij morskih beregovyh gidrometeorologicheskih stancij i postov. – RD 52.10.216-89. [Instructions for sea coastal hydrometeorological stations and posts for hydrological observations data processing and quality control. No. 52.10.216-89]. 1990. Moscow, Gidrometeoizdat, 140 p.

Naumova L.P. 1983. Ocenka vliyaniya propuskov nablyudenij na znacheniya klimaticheskih harakteristik [Estimation of effect of missing values on climatological characteristics]. *Trudy GGO – Proc. Voeikov Main Geophysical Observatory*, vol. 475, pp. 20-25.

Rukovodstvo po specializirovannomu klimaticheskomu obsluzhivaniyu ekonomiki. Pod red. N.V. Kobyshevoj [Guide for special climatological service for economics. Kobysheva (ed.)]. 2008. St-Petersburg, 336 p.

Yaglom A.M. 1963. *Statisticheskie metody ekstrapolyacii meteorologicheskih polej* – V kn.: Tr. Vsesoyuz. nauchn. meteorol. soveshch. «Dinamicheskaya meteorologiya» [Statistical method of extrapolation of meteorological fields – In Proc. All-union sci. meteorol. seminar «Dynamical meteorology»]. Leningrad, Gidrometeoizdat, 280 p.

Bowsher C.G., Swain P.S. 2012. Identifying sources of variation and the flow of information in biochemical networks. – Proc. Natl. Acad. Sci. USA, vol. 109(20), pp. 1320-1329.

MacKay D.J.C. 2003. Information Theory, Inference, and Learning Algorithms. – Cambridge University Press, 628 p.

Massetti L. 2014. Analysis and estimation of the effects of missing values on the calculation of monthly temperature indices. – Theor. Appl. Climatol., vol. 117, iss. 3-4, pp. 511-519.

Tillé Y. 2006. Sampling Algorithms. – Springer Series in Statistics. New York, Springer, 216 p., doi: 10.1007/0-387-34240-0.

Wilks D. 1997. Resampling hypothesis tests for autocorrelated fields. – J. Clim., vol. 10, pp. 65-82.

Yilmaz A.E., Aktas S. 2017. Autocorrelation Corrected Standard Error for Two Sample t-test Under Serial Dependence. – Hacettepe Journal of Mathematics and Statistics, vol. 46(6), pp. 1199-1210, doi: 10.15672/HJMS.201611515847.