

Новый подход к восстановлению пропущенных данных о площади термокарстовых озер Арктики

Попков Ю.С.

Федеральный исследовательский центр "Информатика и управление" РАН, Москва, Россия
popkov.yuri@gmail.com

Мельников А.В.

АУ «Югорский НИИ информационных технологий»,
Ханты-Мансийск, Россия
melnikovav@uriit.ru

Полищук Ю.М.

АУ «Югорский НИИ информационных технологий»,
Ханты-Мансийск, Россия
yupolishchuk@gmail.com

Сокол Е.С.

АУ «Югорский НИИ информационных технологий»,
Ханты-Мансийск, Россия
sokoles@uriit.ru

Полищук В.Ю.

Институт мониторинга климатических и экологических систем СО РАН,
Томск, Россия
liquid_metal@mail.ru

Аннотация. В задачах прогнозирования объемов метана как одного из важных парниковых газов, рассматриваемых в перечне причин современного глобального потепления, необходимо использовать наряду с климатическими характеристиками и данные о динамике площадей термокарстовых озер в зонах мерзлоты, которые обычно получают с помощью космических снимков. Ввиду большого количества пасмурных дней на северных территориях удается получить лишь малое количество безоблачных снимков, что приводит к значительным пропускам во временных рядах площадей озер. Для восстановления пропущенных значений площадей озер в работе предлагается использовать новый подход к восстановлению пропусков, основанный на методах и алгоритмах энтропийно-рандомизированного машинного обучения. Этот подход предполагает проводить восстановление пропущенных значений в экспериментальных данных о площадях термокарстовых озер с использованием временных рядов среднегодовой температуры и годовой суммы осадков. В качестве экспериментальных данных о площадях термокарстовых озер и климатических параметрах (температура и сумма осадков) использованы результаты исследований, проведенных в арктической зоне Западной Сибири в период с 1973 по 2007 г. Исследования проведены на девяти тестовых участках, выбранных в разных зонах мерзлоты (сплошной, прерывистой и островной). Данные о среднегодовой температуре и годовой сумме осадков для каждого тестового участка получены методом реанализа. Разработанный алгоритм восстановления пропущенных значений в рамках указанного подхода реализован с использованием инструментов MATLAB R2019a. Рассчитаны пропущенные значения для выбранных девяти тестовых участков. Для иллюстрации приведены временные ряды значений площади озер, температуры и осадков на одном из тестовых участков. Проведен анализ погрешностей восстановления пропусков, который показал, что разработанный алгоритм позволяет восстанавливать пропущенные значения площадей озер по данным об изменении температуры и осадков с практически приемлемой точностью.

Ключевые слова: рандомизация, машинное обучение, термокарстовые озера, климатические параметры, моделирование, восстановление пропущенных данных.

ВВЕДЕНИЕ

Особую важность рандомизированный подход представляет для решения задач прогнозирования динамики накопления парниковых газов в термокарстовых озерах арктической зоны в связи с их влиянием на глобальные климатические изменения, что может явиться основой разработки и функционирования систем адаптации к меняющимся условиям среды обитания на различных управленческих уровнях. С грядущим глобальным потеплением климата в ближайшие десятилетия будут ускоряться процессы таяния мерзлых пород, приводя к дополнительному высвобождению углекислого газа, а также метана как продукта жизнедеятельности микроорганизмов, перерабатывающих оттаявшее органическое вещество. Это способно внести дополнительный ощутимый вклад в потепление климата, что вызывает озабоченность мировой общественности. Осознание этого послужило поводом для принятия Парижского климатического соглашения (2016), предполагающего разработку в разных странах мер, способных не допустить повышение среднегодовой температуры Земли более чем на 1,5 °C до 2050 года. Разработка таких мер на региональном уровне для арктических регионов невозможна без формирования обоснованных прогнозов объемов эмиссии метана и углекислого газа, которые могут быть основаны на знаниях о пространственно-временной динамике полей озер на территории регионов.

Задачи прогнозирования динамики накопления парниковых газов в термокарстовых озерах на ближайшие десятилетия

тилетия требуют использования данных о временных рядах площадей озер и климатических параметров (температуры, осадков).

Ввиду большой степени заболоченности арктических территорий данные о площадях озер могут быть получены только с использованием спутниковых снимков. Ввиду большого количества пасмурных дней на северных территориях удастся получить безоблачные снимки лишь в отдельные годы. В результате этого полученные временные ряды площадей озер обладают значительным числом пропущенных значений. Вопросы восстановления пропусков во временных рядах, применительно к данным о площадях озер, в настоящее время разработаны недостаточно. Наиболее перспективным подходом к восстановлению пропусков в наших условиях рассматривается использование энтропийно-рандомизированных методов, показавших, согласно [1], высокую эффективность в решении задач глобальной экономики, демографии и др. Однако методические вопросы восстановления пропусков во временных рядах площадей озер в рамках энтропийно-рандомизированного подхода не разработаны, что и явилось целью настоящей работы.

ДАННЫЕ

На рис. 1 приведена схема размещения тестовых участков (ТУ) для проведения исследований, направленных на получение данных о временных рядах площадей термокарстовых озер, среднегодовой температуры и годовой суммы осадков.

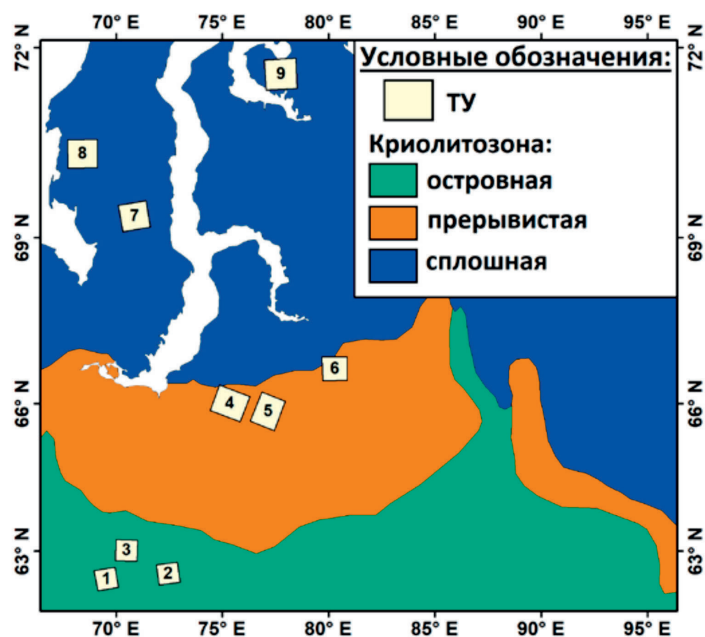


Рис. 1. Схема расположения тестовых участков на исследуемой территории Западной Сибири

Для получения данных о площадях озер использованы космические снимки среднего разрешения (30 м) Landsat, полученные в период с 1973 по 2007 г. Методические вопросы проведения этих исследований изложены в [2]. Данные о климатических характеристиках получены с использованием процедур реанализа данных [2].

Для примера в табл. 1 приведены данные о средней

площади термокарстовых озер $\bar{S}_{\text{изм}}$, среднегодовой температуре $\bar{T}$, годовой сумме осадков $\bar{R}$ на ТУ - 5.

Таблица 1.

Данные о площади, температуре и осадках на ТУ-5.

Годы \ Данные	$\bar{S}_{\text{изм}}$, га	$\bar{T}$, °C	$\bar{R}$, мм	$\bar{S}_{\text{восст}}$, га
1	2	3	4	5
1973	67,90	-1,00	650,30	67,90
1974	-	-3,50	494,25	65,43
1975	-	-2,00	684,80	67,12
1976	-	-2,50	487,30	65,06
1977	-	-2,50	486,50	65,18
1978	-	-3,50	706,40	68,32
1979	-	-2,00	700,10	67,49
1980	-	-2,50	556,50	65,87
1981	-	-1,00	534,60	64,85
1982	-	-2,50	489,95	65,06
1983	-	-1,00	578,80	65,58
1984	68,89	-2,50	465,05	68,89
1985	-	-3,50	482,20	65,40
1986	-	-3,50	577,30	66,68
1987	-	-2,50	467,50	64,78
1988	64,27	-1,00	501,20	64,27
1989	-	-1,00	487,70	64,37
1990	-	-2,50	462,10	64,85
1991	-	-1,00	551,50	65,44
1992	-	-3,50	585,30	66,78
1993	-	-1,00	471,50	64,30
1994	-	-2,50	562,50	65,95
1995	-	0,50	578,90	64,87
1996	-	-1,00	571,35	65,68
1997	-	-2,50	521,10	65,66
1998	-	-3,50	544,80	66,11
1999	-	-2,50	563,15	66,10
2000	-	-1,00	561,30	65,17
2001	62,36	-2,00	430,20	62,36
2002	-	-2,06	520,15	65,21
2003	62,31	-0,48	429,20	62,31
2004	-	-1,80	338,40	62,90
2005	-	0,44	343,90	62,07
2006	-	-2,78	222,00	61,86
2007	66,96	0,14	325,40	66,96

Как видно из табл. 1 (столбец 2), данные о площадях озер имеют большое число пропущенных значений, что является характерным для всех исследованных тестовых участках. Заметим, что в столбце 5 (табл. 1) представлены

восстановленные (модельные) данные о средней площади озер. В табл. 1 вместо модельных данных которых сохранены значения реальных измерений.

Рассмотрим вопросы подготовки данных к расчетам. Данные по площади $\tilde{S}$, температуре $\tilde{T}$, осадкам $\tilde{R}$ собраны по каждому из исследованных девяти тестовых участков. Для расчетов преобразуем данные к стандартному (нормализованному) виду по следующим формулам:

$$\begin{aligned} S &= \frac{\tilde{S} - \tilde{S}_{min}}{\tilde{S}_{max} - \tilde{S}_{min}}, \\ T &= \frac{\tilde{T} - \tilde{T}_{min}}{\tilde{T}_{max} - \tilde{T}_{min}}, \\ R &= \frac{\tilde{R} - \tilde{R}_{min}}{\tilde{R}_{max} - \tilde{R}_{min}}. \end{aligned} \quad (1)$$

где $\tilde{S}$, $\tilde{T}$, $\tilde{R}$ – средняя площадь, среднегодовая температура и сумма годовых осадков в натуральных величинах (га, °C, мм, соответственно), S , T , R – нормализованные значения площади, температуры и осадков (отображенные на промежутке $[0,1]$), нижний индекс у всех показателей означает минимальное и максимальное значение выборки.

МОДЕЛИ И АЛГОРИТМЫ

Технология энтропийно-рандомизированного прогнозирования реализуется в следующей последовательности этапов [1]. Сначала формируется предсказательная рандомизированная параметрическая модель (РПМ), синтезируются ее определяющие параметры и согласуется с ней необходимое информационное обеспечение. РПМ преобразует массив реальных входных данных $X = [x^1, \dots, x^s]$, где $x(j) \in \mathbf{R}^n$ в модельный выход, характеризуемый матрицей $Z = [z(1), \dots, z(s)]$, где $z(j) \in \mathbf{R}^m$. В общем случае это преобразование предполагается динамическим, т.е. модельный выход, наблюдаемый в момент времени j , зависит от входа, наблюдаемого на некотором историческом интервале $j - \varrho, \dots, j$, т.е. от матрицы $X_\varrho(j) = [x^{j-\varrho}, \dots, x^j]$. Математическим выражением этой связи является векторный функционал $\hat{\Omega}(X_\varrho(j)|\alpha, P(\alpha))$ со случайными параметрами α интервального типа

$$\alpha \in \mathcal{A} = [\alpha^-, \alpha^+]. \quad (2)$$

Вероятностные свойства параметров характеризуются функцией плотности распределения вероятности (ПРВ) $P(\alpha)$, которая предполагается непрерывно-дифференцируемой.

Выход РПМ в j -й момент времени (момент измерения) представляет собой ансамбль $\hat{Z}(j | P(\alpha))$ случайных векторов

$$\hat{z}(j|\alpha) = \hat{\Omega}(X_\varrho(j)|\alpha, P(\alpha)), \quad j = \overline{1, s}. \quad (3)$$

Для имитации влияния измерительных ошибок на выходе объекта вводятся случайные шумы $\xi \in \mathbf{R}^m$ интервального типа:

$$\xi^j \in \Xi_j = [\xi_-^j, \xi_+^j], \quad j = \overline{1, s} \quad (4)$$

с непрерывно-дифференцируемыми функциями ПРВ $Q_j(\xi^j)$, $j = \overline{1, s}$, в соответствии с которыми генерируется для каждого момента измерения выхода объекта ансамбль $\mathcal{F}(j|Q_j(\xi^j))$. Набор случайных векторов (шумов измерений) за весь интервал измерений описывается матрицей

$$K = [\xi^{(j)}, j = \overline{1, s}], \quad (5)$$

которая характеризуется совместной функцией ПРВ $Q(K)$. Если шумы в измерениях статистически независимые, то

$$Q(K) = \prod_{j=1}^s Q_j(\xi^{(j)}). \quad (6)$$

Наблюдаемый выход РПМ представим в виде:

$$v(j|\alpha, \xi^{(j)}) = \hat{\Omega}(X_\varrho(j)|\alpha, P(\alpha)), \quad j = \overline{1, s}. \quad (7)$$

Случайные векторы (6) образуют ансамбли, математическое ожидание которых имеет вид:

$$\begin{aligned} \mathcal{M}\{v(j|\alpha, \xi^{(j)})\} &= \int_{\mathcal{A}} \hat{z}(j|\alpha) P(\alpha) d\alpha + \\ &+ \int_{\Xi_j} Q_j(\xi^{(j)}) \xi^{(j)} d\xi^{(j)} = \\ &= \mathcal{W}[P(\alpha), Q_j(\xi^{(j)})], \quad j = \overline{1, s}. \end{aligned} \quad (8)$$

Второй этап технологии рандомизированного прогнозирования связан с обучением РПМ. т.е. с оцениванием ПРВ ее параметров и шумов измерений. Он реализуется с помощью следующего алгоритма [3]:

$$[P^*(\alpha), Q^*(K)] = \arg \max \mathcal{H}[P^*(\alpha), Q^*(K)] \quad (9)$$

на множестве нормированных функций $P^*(\alpha), Q^*(K)$, для которых выполняются условия эмпирических балансов:

$$\mathcal{W}[P(\alpha), Q_j(\xi^{(j)})] = y^{(j)}, \quad j = \overline{1, s} \quad (10)$$

где $y^{(j)} \in \mathbf{R}^m$ - вектор реальных измерений выхода объекта.

Задача (9-10) относится к классу функциональных энтропийно-линейных уравнений ляпуновского типа [4], которые имеют аналитическое решение, получаемое с помощью множителей Лагранжа $\Theta = [\theta^j, j = \overline{1, s}]$ (векторы $\theta^j \in \mathbf{R}^m$):

$$\begin{aligned} P^*(\alpha) &= \frac{\exp(-\sum_{j=1}^s \langle \theta^j, \hat{z}(j|\alpha) \rangle)}{\mathcal{P}(\Theta)}, \\ Q_j^*(\xi^{(j)}) &= \frac{\exp(-\langle \theta^j, \xi^{(j)} \rangle)}{Q_j(\theta^j)}, \quad j = \overline{1, s}; \\ Q_j^*(\xi^{(j)}) &= \frac{\exp(-\langle \theta^j, \xi^{(j)} \rangle)}{Q_j(\theta^j)}, \quad j = \overline{1, s}; \end{aligned}$$

$$Q(K) = \prod_{j=1}^s Q_j^*(\xi^{(j)}). \quad (11)$$

В знаменателях этих выражений стоят нормировочные константы

$$\mathcal{P}(\theta) = \int_{\mathcal{A}} \exp(-\sum_{j=1}^s \langle \theta^j, \hat{z}(j|\alpha) \rangle) d\alpha, \\ Q_j(\theta^j) = \int_{\Xi_j} \exp(-\langle \theta^j, \xi^{(j)} \rangle) d\xi^{(j)}, \quad j = \overline{1, s}. \quad (12)$$

Оптимальные ПРВ и нормировочные константы параметризованы множителями Лагранжа, которые определяются решением следующих балансовых уравнений:

$$\frac{\mathcal{E}(\theta)}{\mathcal{P}(\theta)} + \frac{\mathcal{T}_j(\theta^j)}{Q_j(\theta^j)} = y^{(j)}, \quad j = \overline{1, s} \quad (13)$$

где

$$\mathcal{E}(\theta) = \int_{\mathcal{A}} \hat{z}(j|\alpha) \exp\left(-\sum_{j=1}^s \langle \theta^j, \hat{z}(j|\alpha) \rangle\right) d\alpha, \quad (14) \\ \mathcal{T}_j(\theta^j) = \int_{\Xi_j} \xi^{(j)} \exp(-\langle \theta^j, \xi^{(j)} \rangle) d\xi^{(j)}, \quad j = \overline{1, s}.$$

В нашем случае измеренные данные о площади озер, обладают большим числом пропущенных значений. Для восстановления пропущенных данных воспользуемся принципом энтропийной рандомизации по площади, используя имеющиеся данные по температуре и осадкам. Известно, что на площадь озер влияет температура и осадки, и в первом приближении это влияние можно описать линейной зависимостью с шумом в виде:

$$S[n] = \alpha T[n] + \beta R[n] + \xi[n]. \quad (15)$$

Коэффициенты α, β - случайные, интервальные:

$$\alpha \in \mathcal{A} = [\alpha^-, \alpha^+], \quad \beta \in \mathcal{B} = [\beta^-, \beta^+]. \quad (16)$$

Обозначим ПРВ параметров $P(\alpha), F(\beta)$.

Шум также стандартизованный и интервальный:

$$\xi[n] \in \Xi_j = [\xi^-, \xi^+]. \quad (17)$$

Обозначим ПРВ шума $Q_n(\xi[n])$.

Далее, применяя алгоритм рандомизированного машинного обучения, получим:

$$\mathcal{H} = - \int_{\mathcal{A}} P(\alpha) \ln P(\alpha) d\alpha - \int_{\mathcal{B}} F(\beta) \ln F(\beta) d\beta - \\ - \sum_{m=1}^k \int_{\Xi_m} Q_m(\xi[m]) \ln Q_m(\xi[m]) d\xi[m] \Rightarrow \max \quad (18)$$

при условии нормировки:

$$\int_{\mathcal{A}} P(\alpha) d\alpha = 1, \\ \int_{\mathcal{B}} F(\beta) d\beta = 1, \\ \int_{\Xi_m} Q_m(\xi[m]) d\xi[m] = 1, \quad m = \overline{1, k} \quad (19)$$

и эмпирических балансов:

$$\int_{\mathcal{A}} P(\alpha) \alpha T[m] d\alpha + \int_{\mathcal{B}} F(\beta) \beta R[m] d\beta + \\ + \int_{\Xi_m} Q_m(\xi[m]) \xi[m] d\xi[m] = S[m], \quad m = \overline{1, k}. \quad (20)$$

Решение задачи (18) имеет вид:

$$P^*(\alpha, \theta) = \frac{\exp(-\alpha l_r(\theta))}{\mathcal{P}(\theta)}, \\ F^*(\beta, \theta) = \frac{\exp(-\beta h_r(\theta))}{\mathcal{F}(\theta)}, \quad (21) \\ Q_m^*(\xi[m], \theta) = \frac{\exp(-\theta_m \xi[m])}{Q_j(\theta_m)}, \quad m = \overline{1, k}$$

где $\theta = \{\theta_1, \dots, \theta_k\}$ - множители Лагранжа;
- нормировочные коэффициенты

$$\mathcal{P}(\theta) = \int_{\mathcal{A}} \exp(-\alpha l_r(\theta)) d\alpha; \\ \mathcal{F}(\theta) = \int_{\mathcal{B}} \exp(-\beta h_r(\theta)) d\beta \quad (22) \\ Q_j(\theta_m) = \int_{\Xi} \exp(-\theta_m \xi[m]) d\xi[m], \quad m = \overline{1, k} \\ l_r(\theta) = \sum_{m=1}^k \theta_m T[m], \quad h_r(\theta) = \sum_{m=1}^k \theta_m R[m].$$

Для определения значений множителей Лагранжа нужно решить следующую систему уравнений:

$$L(\theta)T(m) + K(\theta)R(m) + Gm(\theta) = S(m), \quad m = \overline{1, k} \quad (23)$$

$$L(\theta) = \frac{\exp(-\alpha^- l_r(\theta))(\alpha^- l_r(\theta) + 1)}{\exp(-\alpha^- l_r(\theta)) - \exp(-\alpha^+ l_r(\theta))} + \\ + \frac{\exp(\alpha^+ l_r(\theta))(-\alpha^+ l_r(\theta) + 1)}{\exp(\alpha^+ l_r(\theta)) - \exp(-\alpha^+ l_r(\theta) + 1)} \\ K(\theta) = \frac{\exp(-\beta^- h_r(\theta))(\beta^- h_r(\theta) + 1) +}{\exp(-\beta^- h_r(\theta)) - \exp(-\beta^+ h_r(\theta))} + \\ + \frac{\exp(\beta^+ h_r(\theta))(-\beta^+ h_r(\theta) + 1)}{\exp(\beta^+ h_r(\theta)) - \exp(-\beta^+ h_r(\theta) + 1)} \\ Gm(\theta) = \frac{\exp(-\xi^-[m]\theta_m)(\xi^-[m]\theta_m + 1)}{\exp(-\xi^-[m]\theta_m) - \exp(-\xi^+[m]\theta_m)} + \\ + \frac{\exp(-\xi^+[m]\theta_m)(\xi^+[m]\theta_m + 1)}{\exp(-\xi^+[m]\theta_m) - \exp(-\xi^+[m]\theta_m)}. \quad (24)$$

Используя модель для расчета всех пропущенных данных и сэмплируя ПРВ (24), построим ансамбль траекторий $S[n]$. Вычислим среднюю траекторию и по ней заполним недостающие данные.

После этого преобразуем данные от нормализованных величин к натуральным (га). Для этого произведем действие, обратное к (1):

$$\tilde{S}_{\text{восст}} = S * (\tilde{S}_{\text{max}} - \tilde{S}_{\text{min}}) + \tilde{S}_{\text{min}}. \quad (25)$$

РЕЗУЛЬТАТЫ

В соответствии с алгоритмом были восстановлены данные на всех тестовых участках. Для иллюстрации на рис. 2. представлены результаты восстановления данных на тестовом участке № 8.

На рис. 2 (а) изображен график временного хода восстановленных значений площади озер. Сплошными чер-

ными точками отображены реальные (измеренные) значения. Также изображены графики временного хода данных температуры и осадков на рис. 2 (b) и рис. 2 (c) соответственно.

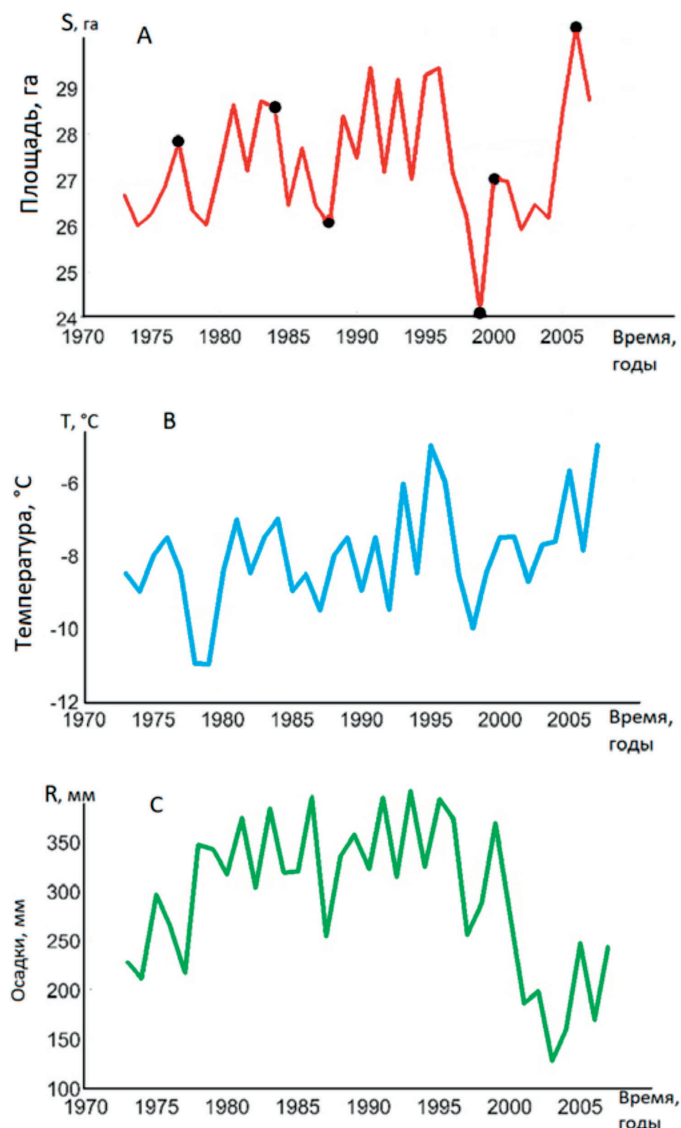


Рис 2. Временные ряды значений восстановленных данных о средней площади (а), результаты измерений средней температуры озер (b) и годовой суммы осадков (с) на ТУ-8

ОЦЕНКА АДЕКВАТНОСТИ МОДЕЛИ

Оценка погрешности моделирования (восстановления) проводилась как оценка среднего отклонения восстановленных значений от измеренных данных.

В табл. 2 приведены на всех тестовых участках измеренные и восстановленные (моделированные) значения в те годы, в которых были проведены измерения площадей озер.

Для каждого тестового участка рассчитаем среднее отклонение восстановленных данных о площади от измеренных значений по следующей формуле:

$$\Delta_s = \frac{1}{m} \sum_{i=1}^m \frac{|\tilde{S}_{\text{восст}i} - \tilde{S}_{\text{изм}i}|}{\tilde{S}_{\text{изм}i}}, \quad (27)$$

где $\tilde{S}_{\text{изм}i}$ – значение средней площади, полученное в результате измерений, $\tilde{S}_{\text{восст}i}$ – восстановленные значения средней площади, m – количество моделируемых точек на конкретном ТУ.

Таблица 2.

Восстановленные и измеренные данные на всех ТУ.

ТУ	Годы	$\tilde{S}_{\text{изм}}$, га	$\tilde{S}_{\text{восст}}$, га
5	1973	67,90	66,61
	1984	68,89	64,86
	1988	64,27	64,77
	2001	62,36	64,11
	2003	62,31	63,5
	2007	66,96	62,12
6	1973	47,60	42,95
	1987	37,05	36,68
	1988	39,43	34,47
	2001	33,51	37,46
	2003	34,30	35,79
	2007	48,66	33,03
7	1973	19,62	19,53
	1984	15,64	17,78
	1988	19,69	18,33
	2001	18,52	17,22
	2003	16,41	17,47
	2007	19,68	17,59
14	1984	23,04	21,7
	1989	23,07	21,51
	1999	21,37	22,48
	2001	20,94	21,71
	2006	21,21	20,5
	2007	20,54	20,85
15	1973	28,65	30,53
	1987	32,25	30,32
	2001	29,93	29,65
	2006	26,92	26,69
16	1973	17,45	16,58
	1987	16,96	15,36
	2001	14,58	15,85
	2006	16,09	15,26
	2007	16,53	16,1
20	1988	18,38	17,49
	1999	16,08	17,25
	2000	17,02	17,32

	2001	17,08	16,97
	2006	17,78	16,85
22	1977	27,82	26,36
	1984	28,59	28,32
	1988	26,02	26,66
	1999	24,14	27,8
	2000	27,04	27,7
	2006	30,32	26,34
24	1977	25,62	22,94
	1981	22,59	22,86
	1988	24,24	23,73
	1999	23,81	23,48
	2001	22,96	23,24
	2006	23,76	22,86

Полученные данные приведены в табл. 3.

Таблица 3.

Погрешности восстановления данных

ТУ	Δ_s
1	0,04
2	0,15
3	0,08
4	0,04
5	0,04
6	0,06
7	0,04
8	0,07
9	0,04
Среднее значение	0,05

Для оценки погрешности восстановления были рассчитаны величины среднего отклонения модельных значений площади от измеренных величин на всех тестовых участках по формуле:

$$\Delta = \frac{1}{k} \sum_{i=1}^k \frac{|\tilde{S}_{\text{восст}i} - \tilde{S}_{\text{изм}i}|}{\tilde{S}_{\text{изм}i}}, \quad (28)$$

где $\tilde{S}_{\text{изм}i}$ – измеренные значения площади, $\tilde{S}_{\text{восст}i}$ – восстановленные значения средней площади, k – количество всех измеренных значений средней площади по всем тестовым участкам.

В нашем случае было получено следующее значение: $\Delta = 0,06$, при этом среднеквадратичное отклонение, рассчитанное по стандартной формуле, равно 0,09. Следовательно, погрешность восстановления пропусков по разработанному алгоритму не превышает в среднеквадратическом 9%, что можно рассматривать как практически приемлемый результат.

ЗАКЛЮЧЕНИЕ

В статье рассмотрены методические вопросы нового подхода к восстановлению пропущенных значений в экспериментальных данных о площадях термокарстовых озер с использованием временных рядов среднегодовой температуры и годовой суммы осадков. В качестве экспериментальных данных о площадях термокарстовых озер и климатических параметрах (температура и сумма осадков) использованы результаты исследований, проведенных в арктической зоне Западной Сибири в период с 1973 по 2007 г. Для проведения исследований выбраны девять тестовых участков, по три в каждой из зон мерзлоты (сплошной, прерывистой и островной). Данные о среднегодовой температуре и годовой сумме осадков для каждого тестового участка получены методом реанализа. Данные о площадях озер являются результатами дистанционных измерений по космическим снимкам. Ввиду большого количества пасмурных дней на севере Западной Сибири в указанный период удалось получить не более восьми значений, полученных в те годы, когда имелись безоблачные снимки. Поэтому для восстановления пропущенных значений временного ряда, число которых значительно превышает число имеющихся измеренных значений, наиболее перспективным оказался подход, основанный на энтропийно-рандомизированном моделировании, методология которого разработана в [1, 3].

Разработан алгоритм восстановления пропущенных значений в рамках указанного подхода, реализованный с использованием инструментов MATLAB R2019a. Рассчитанные пропущенные значения для указанных девяти тестовых участков представлены в табличном виде. Для иллюстрации приведены временные ряды значений площади озер, температуры и осадков на одном из тестовых участков. Проведенный анализ погрешностей восстановления пропусков показал, что разработанный алгоритм позволяет восстанавливать пропущенные значения площадей озер по данным об изменении температуры и осадков с практически приемлемой точностью.

БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке гранта РФФИ по проекту № 19-07-00282.

ЛИТЕРАТУРА

1. Попков Ю.С. Рандомизированное машинное обучение при ограниченных объемах данных. / А.Ю. Попков, Ю.А. Дубнов. – Москва: УРСС, 2019. – 310 с.
2. Полищук В.Ю. Геоимитационное моделирование полей термокарстовых озер в зонах мерзлоты / В.Ю. Полищук, Ю.М. Полищук. – Ханты-Мансийск: УИП ЮГУ, 2013. – 129 с.
3. Popkov Y.S. New Method of Entropy-Robust Estimation for Randomized Models under Limited Data / Y.S. Popkov, A.Y. Popkov // Entropy. – 2014. – vol. 16. – P. 675-698.
4. Иоффе А.Д., Теория экстремальных задач / А.Д. Иоффе, В. Тихомиров. – Москва: Наука, 1974. – 435 с.
5. Hastie T. The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. Friedman // Springer. – 2001. – 533 p.
6. Bishop C.M. Pattern Recognition and Machine Learning / C.M. Bishop, M. Jordan, J. Kleinberg // Springer, Series: Information Theory and Statistics. – 2006. – 345 p.

7. Tarasov A. Forecasting US money growth using economic uncertainty measures and regularization techniques // *International Journal of Forecasting*. – 2019. – vol. 35. – P. 443-457.
8. Marcellino M. A comparison of direct and iterated multisteps AR methods for forecasting macroeconomic time series / M. Marcellino, J.H. Stock, M. Watson // *Journal of Econometrics*. – 2006. – vol. 135, No. 1-2. – P. 499-526.
9. Eitrheim O. Testing the adequacy of smooth transition autoregressive models / O. Eitrheim, T. Terasirta // *Journal of Econometrics*. – 1996. – vol. 74, No. 1-2. – P. 59-75.
10. Molodtsova T. Out-of-sample exchange rate predictability with TR fundamentals / T. Molodtsova, D. Papell // *Journal of International Economy*. – 2009. – vol. 77. – P. 167-180.
11. Wang R. Forecasting the exchange rate using nonlinear Taylor rule based model / R. Wang, B. Motley, M. Stamatogiannis // *International Journal of Forecasting*. – 2019. – vol. 35. – P. 429-442.
12. Canale A., Bayesian nonparametric forecasting of nonmonotonic functional time series / A. Canale, M. Ruggiero // *Electronic Journal of Statistics*. – 2016. – vol. 10, No. 2. – P. 3265-3288.
13. Дубнов Ю.А., Байесовская идентификация параметров смеси нормальных распределений / Ю.А. Дубнов, А.В. Булычев // *Информационные технологии и вычислительные системы*. – 2017. – № 1. – С. 101-114.
14. Frazier D.T. Approximate Bayesian forecasting / D.T. Frazier, W. Maneesoontherm, G.M. Martin, B.P.M. McCabe // *International Journal of Forecasting*. – 2019. – vol. 35. – P. 521-539.
15. Beaumont M.A. Approximate Bayesian computation in population genetics / M.A. Beaumont, W. Zhang, D.J. Baldung // *Genetics*. – 2001. – vol. 162, No. 4. – P. 2025-2035.
16. McAdam P. Euro-area real time density forecasting with financial or labour market friction. / P. McAdam, A. Warne // *International Journal of forecasting*. – 2019. – vol. 35. – P. 580-600.
17. Alkema L. The UN probabilistic population projection: an introduction to demographic forecasting with uncertainty / Alkema L., Geriand P., Raftery A. Wilmoth T. // *Forecasting*. – 2015. – vol. 37. – P. 19-24.
18. Blum M.G.B. A comparison review of dimension reduction methods in approximate Bayesian computation / M.G.B. Blum, M.A. Nunes, D. Prangle, S.A. Sisson // *Statistical Science*. – 2013. – vol. 28, No. 2. – P. 189-208.
19. Воронцов К. В. Математические методы обучения по прецедентам. – Курс лекций. – Москва: МФТИ, 2013. – 245 с.
20. Цыпкин Я.З. Основы теории обучающихся систем. – Москва: Наука, 1970. – 398 с.
21. Вапник В.Н. Теория распознавания образов. / В.Н. Вапник, А.Я. Червоненкис. – Москва: Наука, 1974. – 237 с.
22. Witten J.H. Data Mining: Practical Machine Learning Tools and Techniques / J.H. Witten, E. Frank // *Morgan Kaufmann*. – 2005. – 458 p.
23. Gaillard P. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting / P. Gaillard, K. Goude, R. Nedellec // *International Journal of Forecasting*. – 2016. – vol. 32. – P. 1038-1050.
24. Mangalova E. Wind power forecasting using the k-nearest neighbours algorithm / E. Mangalova, E. Agafonov // *International Journal of Forecasting*. – 2014. – vol. 30. – P. 402-406.
25. Zhang K. K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting / K. Zhang, J. Wang // *International Journal of Forecasting*. – 2016. – vol. 32. – P. 1074-1080.
26. Armstrong J. Long-range forecasting. – New York: Wiley, 1985. – 432 p.
27. Wheatcroft E. Interpreting the skill score from forecast performance metrics // *International Journal of Forecasting*. – 2019. – vol. 35. – P. 573-579.
28. Brier C. W. Verification of forecast expressed in term probability // *Monthly Weather Review*. – 1950. – vol. 78, No. 1-3. – P. 25-34.
29. Brocker J. From ensembles forecasts to predictive distribution functions / J. Brocker, L.A. Smith // *Tellus A*. – 2008. – vol. 60, No. 4. – P. 663-678.
30. Christensen H.M. Evaluation of ensemble forecast uncertainty using a new proper score: application to medium-range and seasonal forecast / H.M. Christensen, I.M. Moroz, T.N. Palmer // *Quarterly Journal of the Royal Meteorological Society*. – 2015. – vol. 141, No. 687. – P. 538-549.
31. Lahiri K. Evaluating probability forecasts for GDP declines using alternative methodological / K. Lahiri, J.C. Wang // *International Journal of Forecasting*. – 2013. – vol. 29, No. 1. – P. 175-190.
32. Gneiting T. Probabilistic Forecasting / T. Gneiting, M. Katzfuss // *Annual Review of Statistics and Its Applications*. – 2014. – vol. 1. – P. 125-151.
33. Vidyasagar M. Statistical Learning Theory and Randomized Algorithms for Control // *IEEE Control System Magazine*. – 1998. – vol. 1, No. 17. – P. 69-88.
34. Гранич О.Н. Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах. / О.Н. Гранич, Б.Т. Поляк // *Москва: Наука*, 2002. – 291 с.
35. Biondo A.E. Are random trading strategies more successful than technical ones? / A.E. Biondo, A. Pluchino, A. Rapisarda, D. Helbing // *PLoS ONE*. – 2013. – vol. 6, No. 7, P. 34-45.
36. Lutz W. The end of world population growth. / W. Lutz, S. Sandersen, S. Scherbov // *Nature*. – 2001. – vol. 412, No. 6846. – P. 543-545.
37. Цирлин А.М. Методы усредненной оптимизации и их применение. – Москва: Физматлит, 1997. – 304 с.
38. Shannon C. Communication Theory of Secrecy Systems // *Bell System Technical Journal*. – 1949. – vol. 28, No 4. – P. 656-715.
39. Jaynes E.T. Information Theory and Statistical Mechanics // *Physics Review*. – 1957. – vol. 106. – P. 620-630.
40. Jaynes E.T. Papers on probability, statistics and statistical physics. – Dordrecht: Kluwer Academic Publisher, 1989, 436 p.
41. Jaynes E.T. Probability Theory. The logic and science // *Cambridge: Cambridge University Press*, 2003. – 341 p.

A new approach to recovering missing data on the area of thermokarst lakes in the Arctic

Yuri S. Popkov

Federal Research Center "Informatics and Control" RAS,
Moscow, Russia
popkov.yuri@gmail.com

Andrey V. Melnikov

Ugra Research Institute of Information Technologies,
Khanty-Mansiysk, Russia
melnikovav@uriit.ru

Yury M. Polishchuk

Ugra Research Institute of Information Technologies,
Khanty-Mansiysk, Russia
yupolishchuk@gmail.com

Evgeny S. Sokol

Ugra Research Institute of Information Technologies,
Khanty-Mansiysk, Russia
eugen137@gmail.com

Vladimir Y. Polishchuk

Institute of Monitoring of Climatic and Ecological Systems,
Tomsk, Russia,
liquid_metal@mail.ru

Abstract. In the tasks of forecasting the volumes of methane as one of the important greenhouse gases considered in the list of causes of global warming today, it is necessary to use, along with climatic characteristics, data on the dynamics of the areas of thermokarst lakes in permafrost, which are usually obtained using satellite imagery. Because of a large number of cloudy days in the northern territories allows you to get only a small number of cloudless images, which leads to significant number of missing data in the time series of lakes areas. To restore missing values, it is necessary to use a new approach to recovering gaps based on methods and algorithms of entropy-randomized machine learning. This approach involves the restoration of missing values in the experimental data on the sizes of thermokarst lakes using time series of average annual temperature and annual precipitation. As experimental data on the thermokarst lakes areas and climatic parameters (temperature and amount of precipitation), we used the results of studies conducted in the Arctic zone of Western Siberia from 1973 to 2007. Studies were conducted in

nine test sites selected in different permafrost zones (continuous, discontinuous and insular). Data on the average annual temperature and annual precipitation for each test site were obtained by reanalysis. The developed algorithm for recovering missing values within the framework of this approach is implemented using the MATLAB R2019a tools. The missing values are calculated for the selected nine test sites. To illustrate, the time series of the values of the area of lakes, temperature and precipitation in one of the test sites are shown. An analysis of the omissions recovery errors was carried out, which showed that the developed algorithm allows us to restore the missing values of the lake areas from the data on changes in temperature and precipitation with practically acceptable accuracy.

Keywords: *randomization, machine learning, thermokarst lakes, climatic parameters, modeling, restoration of missing data.*