

# Применение алгоритма восполнения матриц для сжатия и обработки данных о температуре поверхности моря с разреженными ошибками измерений

Шелопут Т. О.<sup>1,2\*</sup>, Петров С. В.<sup>1</sup>, Косолапов И. А.<sup>2</sup>

<sup>1\*</sup>ИВМ РАН, ул. Губкина, 8, Москва, 119333, Россия.

<sup>2</sup>Физтех-школа прикладной математики и информатики, МФТИ, Институтский переулок, 9, Долгопрудный, 141701, Московская область, Россия.

\*Автор(ы), ответственный(ые) за переписку. E-mail(s):  
sheloput@phystech.edu;

Соавторы: spetrov.msk@gmail.com; iakosolapov1@gmail.com;

## Аннотация

Известно, что из-за необходимости хранения и передачи все большего объема геофизических данных актуальной является проблема разработки эффективных алгоритмов сжатия. Для использования геофизических данных в практических приложениях проводится предварительная обработка (например, фильтрация аномалий и интерполяция данных). В данной статье метод, основанный на матричных аппроксимациях, применяется к задачам сжатия, предварительной обработки и интерполяции температуры поверхности моря (ТПМ). Используемый метод основан на аппроксимации матрицы данных как суммы матрицы низкого ранга и разреженной матрицы. Показано, что метод позволяет сжимать с потерями данные о температуре морской поверхности, полученные со спутников, а также может использоваться для заполнения относительно небольших пробелов в данных. Количество действительных чисел, необходимое для восстановления аппроксимации исходного поля, зависит от допустимой ошибки аппроксимации. Для вычисления одного элемента матричной аппроксимации требуются  $R$  умножений, где  $R$  обозначает ранг матрицы. Разреженный компонент матричной аппроксимации содержит элементы, которые алгоритм считает аномалиями, и они, как правило, локализируются вблизи побережья. Возможные причины – фактическое наличие аномалий в данных в прибрежных зонах и специфика формирования исходной матрицы из данных, которая может потребовать корректировки. Следует отметить, что необходимы дальнейшие численные

исследования для формулирования рекомендаций по использованию метода для сжатия, обработки и интерполяции данных.

**Ключевые слова:** Сжатие геофизических данных, Температура поверхности моря, Интерполяция данных, Обработка спутниковых данных

Получено редакцией 01.12.2023; внесены авторские правки 02.09.2024; принята к публикации 27.09.2024

## 1 Введение

Одним из популярных продуктов, предоставляемых сервисами-агрегаторами геофизических данных [1], является реанализ данных наблюдений, полученных с помощью математического моделирования гидротермодинамики морских акваторий. Реанализ, полученный по моделям Мирового океана и другим моделям крупных морских акваторий с детальным пространственно-временным разрешением, занимает относительно большой объем дискового пространства, поэтому возникают трудности с хранением и передачей данных по сети. В связи с этим имеется необходимость в разработке и исследовании эффективности алгоритмов сжатия, учитывающих особенности геофизических данных. Другой актуальной проблемой является возрастающий объем хранимых данных дистанционного зондирования земли (ДЗЗ). Для использования данных ДЗЗ в практических приложениях проводится предварительная обработка, включающая процедуру фильтрации аномалий. Также иногда требуется интерполяция данных в ненаблюдаемые зоны.

В настоящее время существует множество алгоритмов сжатия данных, однако специальных алгоритмов компрессии геофизических данных, учитывающих их структуру и особенности, предложено не так много. Например, данные реанализа представляют собой четырехмерные массивы температуры, солёности, скоростей, в которых значения меняются относительно «гладко». То есть, если представить, что каждый элемент массива является значением функции в узле пространственно-временной сетки математической модели, то можно найти подобласти, в которых изменение значений функции между соседними узлами сетки ограничено и мало по сравнению с самими значениями функции. Известные подходы, связанные со сжатием битовых представлений (алгоритм Хаффмана и другие алгоритмы кодирования битовых последовательностей) не являются достаточно эффективными в данном случае (степень сжатия около 3:1). Более перспективными являются алгоритмы, основанные на вейвлет-преобразовании с последующим квантованием коэффициентов [2], или тензорные разложения [3].

Важным шагом при обработке данных со спутников является коррекция или исключение “ложных” измерений и выбросов. Методы, которые

при этом используются, основаны не только на математической статистике, но и на некоторых эмпирических и инженерных соображениях [4]. Так, например, каждый продукт SST на сайте [1] сопровождается указанием качества измерения в каждой точке по шкале от “0” до “5”, где “0” – необработанные данные или нет измерений, “1” – “облачно”, “2” – данные плохого качества, “3”, “4” – приемлемого качества, “5” – отличного качества. Для получения индекса качества используется целый набор характеристик, учитывающих степень отклонения от среднеклиматических данных, облачность, аэрозоли, наличие больших отклонений между соседними измерениями и вероятность появления льда. Оценка качества полученного продукта проводится при сравнении полученного результата с данными контактных измерений.

Для интерполяции, вероятностного анализа и обработки геофизических данных, как правило, используются подходы, основанные на математической статистике. Соответствующую область науки принято называть геостатистикой. К таким подходам относится известный метод оптимальной интерполяции [5] и кригинг [6]. Статистический анализ данных позволяет также определить наличие ошибок и выбросов в данных. Помимо статистических интерполяторов в геофизике могут применяться также детерминистические интерполяторы (линейная и полиномиальная интерполяция, метод обратных взвешенных расстояний и др.) и алгоритмы, основанные на обучении (нейронные сети).

В настоящей работе рассматривается возможность применения алгоритмов восполнения матриц малого ранга к обработке геофизических данных. В частности, такие алгоритмы предлагается применить к задачам сжатия и интерполяции данных, которые могут содержать разреженные ошибки измерений, или аномалии. Рассматриваемые данные при этом соответствуют измерениям температуры поверхности Черного моря, полученным спутниками Aqua, Terra, Suomi NPP, а также результатам реанализа данных с использованием модели гидротермодинамики Черного моря на базе INMOM [7].

## 2 Описание используемого подхода

Как правило, задачи о восполнении матриц заключаются в поиске всех элементов матрицы по небольшому их числу, в предположении, что искомая матрица имеет малый ранг. Существуют различные способы формальной постановки задачи о восполнении матриц. Рассмотрим следующую постановку.

Пусть дана маска индексов  $\Theta \in \mathbb{N} \times \mathbb{N}$ , в которых заданы значения матрицы  $X \in \mathbb{R}^{n \times n}$ . Рассмотрим оператор  $A_\Theta : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ , такой, что

$$A_\Theta(X)_{ij} = \begin{cases} \frac{X_{ij}}{\sqrt{\rho}}, & (i, j) \in \Theta, \\ 0, & (i, j) \notin \Theta, \end{cases} \quad (1)$$

где  $\rho = \#\Theta/n^2$  – доля известных элементов,  $\#\Theta$  – мощность множества  $\Theta$ .

Задача о восполнении матрицы формулируется следующим образом: найти  $Y \in \mathbb{R}^{n \times n}$ , такую, что

$$Y = \arg \min_{\tilde{X} \in \mathbb{R}^{n \times n}} \|A_\Theta(\tilde{X}) - A_\Theta(X)\|,$$

$$\text{rank}(\tilde{X}) \leq r,$$

где  $r < n$  – предопределенная константа. Значения решения задачи  $X$ , не соответствующие позициям из множества  $\Theta$ , могут быть использованы в целях интерполяции. Метод проекции градиента, введенный в [8] ('Singular Value Projection'), итерация которого описывается как

$$X_{k+1} = \mathcal{P}_r(X_k + \tau A_\Theta^* A_\Theta(X - X_k)),$$

имеет теоретический результат о геометрической сходимости при условии, что неизвестное  $X$  и ошибки итераций  $X_k - X$  сохраняют “неразреженность”, формально определенную как некогерентность в [8]. Оператор  $\mathcal{P}_r$  обозначает оптимальное приближение входной матрицы с рангом не выше  $r$ , которое можно получить с помощью частичного сингулярного разложения. В статье [9] результат о сходимости расширен на случай применения неточных проекций: показано, что геометрическая сходимость сохраняется даже при использовании быстрых приближенных алгоритмов поиска частичного сингулярного разложения.

Более того, метод проекции градиента, с небольшими модификациями, может быть использован для очистки малоранговых данных от разреженных ошибок. Для этого в [10] предложено выбрать два подмножества  $\Theta_a \in \Theta$ ,  $\Theta_b \in \Theta$ , запустить два шага проекции градиента

$$X_{k+1,a} = \mathcal{P}_{r+p}(X_k + \tau A_{\Theta_a}^* A_{\Theta_a}(X - X_k)),$$

$$X_{k+1,b} = \mathcal{P}_{r+p}(X_k + \tau A_{\Theta_b}^* A_{\Theta_b}(X - X_k))$$

с расширенным рангом  $r+p$ , где  $p \approx \log n$ , а затем выбрать  $X_{k+1}$  на основе наиболее сонаправленных подпространств между подпространствами строк и столбцов  $X_{k+1,a}$ ,  $X_{k+1,b}$ . Показано [10], что такая процедура устойчива к разреженным ошибкам и может быть эффективно использована для обнаружения позиций аномальных данных.

### 3 Описание исходного набора данных

Для численных экспериментов по анализу эффективности и применимости алгоритма восполнения матриц для интерполяции, обработки и сжатия были использованы данные о температуре поверхности моря (ТПМ) – данные со спутников и данные реанализа.

Данные о ТПМ со спутников представляют собой информацию о температуре верхнего слоя воды океанов и морей, которая получается после обработки данных со спутниковых приборов. В работе рассматриваются данные со спутников Aqua, Terra, Sentinel и Suomi NPP, пересчитанные на регулярную широтно-долготную сетку  $800 \times 1600$  (800 отсчетов по широте с шагом  $0.01^\circ$ , 1600 отсчетов по долготе с шагом  $0.01^\circ$ ), покрывающую акватории Черного, Азовского и Мраморного морей. Спутники Terra и Aqua оборудованы приборами MODIS (Moderate Resolution Imaging Spectroradiometer), которые имеют 36 спектральных каналов в видимом, ближнем, среднем и тепловом инфракрасном диапазонах и активно используются для мониторинга температуры поверхности океанов, суши и ледяного покрова. Спутник Suomi NPP оборудован инфракрасным радиометром, который используется для измерения температуры поверхности моря и суши на высоких широтах. На борту спутника Sentinel-3 температурные измерения производятся с помощью двух инструментов: Sea and Land Surface Temperature Radiometer (SLSTR) — радиометра, оборудованного инфракрасными камерами, которые позволяют с высокой точностью измерять инфракрасное излучение, излучаемое поверхностью океана и морей; и Ocean and Land Colour Instrument (OLCI) — прибора для измерения цвета океана и суши, который также может предоставлять данные о температуре поверхности океана.

Данные наблюдений со спутников являются моментальными снимками, которые могут не покрывать всю исследуемую акваторию, более того, в некоторых условиях часть данных не наблюдается из-за облаков или исключается на этапе обработки из-за плохого качества. Такие данные могут содержать выбросы и изолированные измерения (во всех соседних узлах сетки значения неопределены).

Данные реанализа получены с использованием вариационной ассимиляции данных наблюдений о ТПМ [11] с перечисленных выше спутников в модели гидрогермодинамики Черного и Азовского морей, основанной на модели INMOM [7]. В численных экспериментах использовались данные о температуре приповерхностного слоя толщиной 1 метр на прямоугольной сетке  $200 \times 306$  (200 отсчетов по широте с шагом  $0.036^\circ$ , 306 отсчетов по долготе с шагом  $0.05^\circ$ ).

Для каждого типа данных из перечисленных выше можно сформировать матрицу  $\hat{T}$  со значениями температуры в узлах сетки. Положение точки на поверхности моря  $\Omega$  можно описать с использованием сферической системы координат  $(\lambda, \phi, r)$ , а именно – парой координат  $\lambda, \phi$ , где

$\lambda$  – долгота,  $\phi$  – широта. Пусть по широте и долготе введены дискретизации  $\lambda_1 < \lambda_2 < \dots < \lambda_N$ ,  $\phi_1 < \phi_2 < \dots < \phi_M$ , такие, что для любой точки  $\vec{x} = (\lambda, \phi, r(\lambda, \phi))$ ,  $\vec{x} \in \Omega$ , выполнено  $\lambda_1 \leq \lambda \leq \lambda_N$ ,  $\phi_1 \leq \phi \leq \phi_M$ . Будем считать, что эти дискретизации соответствуют исходной широтно-долготной сетке, то есть измерение температуры  $T_{ij}$  аппроксимирует температуру в точке с координатами  $(\lambda_i, \phi_j)$  на поверхности моря. Обозначим через  $V$  множество пар индексов  $(i, j)$ , для которых имеются измерения температуры в исходном наборе данных, а через  $\Omega_h$  – множество индексов  $(i, j) \in V$ , таких, что  $(\lambda_i, \phi_j, r(\lambda_i, \phi_j)) \in \Omega$ . Сформируем матрицу данных  $\hat{T} \in \mathbb{R}^{M \times N}$  по следующему правилу:

$$(\hat{T})_{ij} = \begin{cases} T_{ij}, & (i, j) \in \Omega_h, \\ 0, & (i, j) \notin \Omega_h, \end{cases} \quad (2)$$

где индекс  $i$  пробегает все целые значения от 1 до  $N$ , а  $j$  – от 1 до  $M$ . Для матрицы  $\hat{T}$  можно применить алгоритм восполнения, предложенный в работе [10], и получить аппроксимацию этой матрицы вида  $\tilde{T} = T_r + S$ , где  $T_r$  – матрица ранга не выше  $r$ ,  $S$  – разреженная матрица.

## 4 Результаты численных экспериментов.

Для оценки результатов численных экспериментов были введены следующие величины:

1. Ошибка аппроксимации  $E_c = \max_{(i,j) \in \Omega_h} |T_{ij} - \tilde{T}_{ij}|$  – наибольшее по модулю значение отклонения аппроксимации от исходной матрицы.
2. Относительная ошибка аппроксимации в норме Фробениуса  $E_f = \frac{\sum_{(i,j) \in \Omega_h} |T_{ij} - \tilde{T}_{ij}|^2}{\sum_{(i,j) \in \Omega_h} |T_{ij}|^2}$ .
3. Среднеквадратическая ошибка  $MSE = \frac{\sum_{(i,j) \in \Omega_h} |T_{ij} - \tilde{T}_{ij}|^2}{(\#\Omega_h)}$ , где  $\#\Omega_h$  – мощность множества  $\Omega_h$ .

Рассмотрим результаты численного эксперимента по сжатию данных о ТПМ со спутников. Для оценки характеристики сжатия будем использовать две величины:

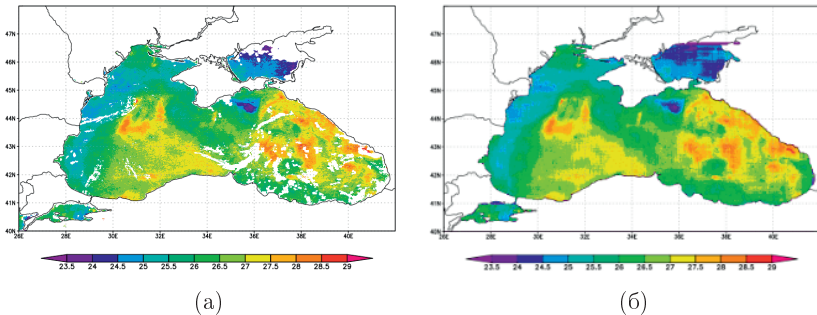
1. Доля хранимых вещественных чисел ( $N_{re}$ ) – отношение количества вещественных чисел, необходимых для вычисления аппроксимации, к количеству исходных значений.
2. Коэффициент сжатия ( $K$ ) – отношение размера исходного к размеру сжатого файла.

В качестве примера рассмотрим результаты реализации алгоритма сжатия ТПМ по данным со спутника Aqua за 2 сентября 2018 г., 11:05.

Общее число ненулевых элементов матрицы  $\hat{T}$ , построенной по этому набору данных, составляет 450604. В Таблице 1 представлена зависимость ошибок аппроксимации данных от характеристик сжатия для ранга 10.

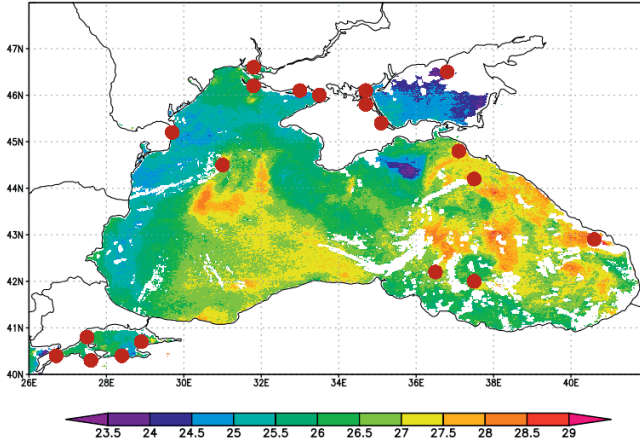
Число разреженных эл.	$N_{re}, \%$	$E_f$	$E_c, ^\circ C$
0	5.33	$1.62 \cdot 10^{-2}$	2.605
3200	6.04	$9.62 \cdot 10^{-3}$	1.098
6400	6.75	$9.13 \cdot 10^{-3}$	0.845
9600	7.46	$8.78 \cdot 10^{-3}$	0.737
14400	8.52	$8.36 \cdot 10^{-3}$	0.644
18358	9.40	$8.05 \cdot 10^{-3}$	0.589
22449	10.31	$7.78 \cdot 10^{-3}$	0.550
29595	11.89	$7.37 \cdot 10^{-3}$	0.500
39130	14.01	$6.89 \cdot 10^{-3}$	0.450

**Таблица 1:** Зависимость ошибок аппроксимации от характеристик сжатия (ранг 10) и количества ненулевых элементов в разреженной матрице  $\hat{S}$  для данных со спутника Aqua за 2 сентября 2018 года, 11:05



**Рис. 1:** Температура поверхности моря по данным со спутника Aqua за 2 сентября 2018 года, 11:05: (а) – исходные данные ЦКП «ИКИ - мониторинг»; (б) – результат применения алгоритма восполнения с 39130 ненулевыми элементами разреженной матрицы

Исходные данные (см. Рис. 1а) можно сравнить с результатом сжатия, соответствующим последней строке Таблицы 1 (см. Рис. 1б). Из рисунков видно, что алгоритм позволяет восполнить пробелы в спутниковых данных. Было исследовано пространственное распределение значений разреженной матрицы.

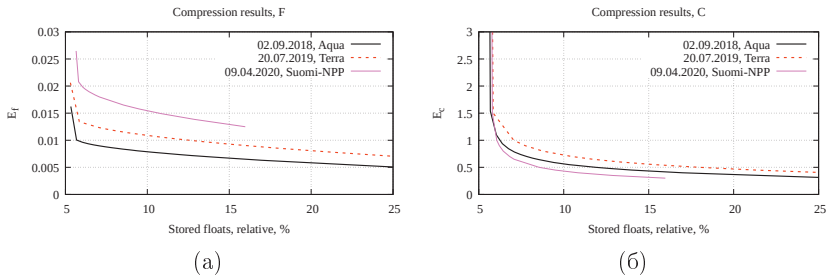


**Рис. 2:** Температура поверхности моря по данным со спутника Aqua за 2 сентября 2018 года, 11:05. Круглыми маркерами отмечено расположение элементов разреженной матрицы, значения которых превышали по модулю  $1,5^{\circ}\text{C}$ .

На Рис. 2 круглыми маркерами отмечено расположение элементов разреженной матрицы, значения которых превышали по модулю  $1,5^{\circ}\text{C}$ . Расположение маркеров можно интерпретировать как расположение «выбросов», поскольку индексы больших по модулю элементов разреженной матрицы характеризуют положения элементов исходного набора данных, наиболее отклоняющихся от его малоранговой («структурированной») составляющей. Однако, можно заметить, что многие маркеры расположены вблизи берегов. Причиной этого явления может быть не только искажение измерений спутников вблизи берегов, но и особенности формирования исходной матрицы данных, а именно, заполнение нулями береговой и ненаблюдаемой зон.

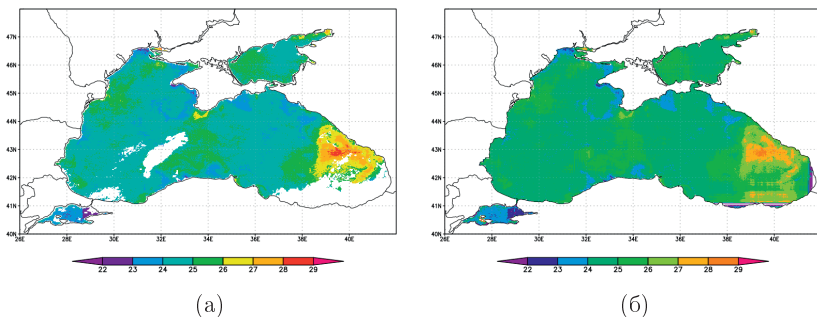
На Рис. 3 для сравнения представлены графики зависимости ошибок аппроксимации  $E_f$  (Рис. 3а),  $E_c$  (Рис. 3б) от характеристик сжатия для разных наборов данных: синей сплошной линией представлены данные Aqua (02.09.2018, 11:05), красным пунктиром – Terra (20.07.2019, 08:30),





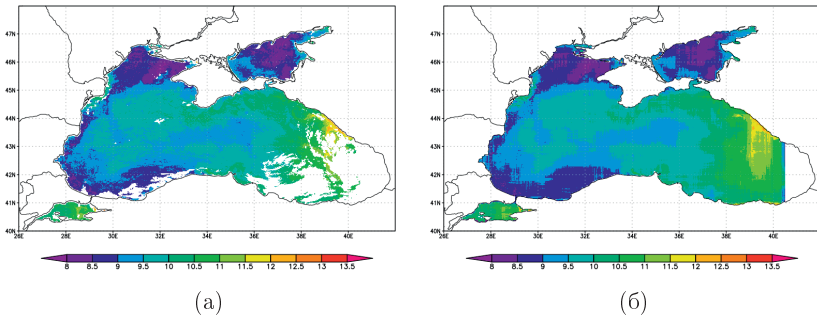
**Рис. 3:** Зависимость ошибки аппроксимации, (а) –  $E_f$ , (б) –  $E_c$  (°C), от  $N_{re}$  (%) для данных со спутников Aqua (02.09.2018, 11:05), Terra (20.07.2019, 08:30), Suomi-NPP (09.04.2020, 23:24)

сиреневой сплошной линией – Suomi-NPP (09.04.2020, 23:24). Как видно из рисунков, характеристики сжатия зависят от исходных данных, однако для всех перечисленных наборов данных при значениях  $N_{re}$  от 10% до 15% ошибка аппроксимации  $E_c$  составляет около 0,5°C и сравнима с ошибкой исходного продукта [12]. Такую аппроксимацию можно считать приемлемой, например, для использования в задачах вариационной ассимиляции. Данные со спутника Terra о ТПМ 20 июля 2019 года, 08:30, представленные на Рис. 4а, можно сравнить с результатом применения алгоритма восполнения с рангом 10 и 57456 ненулевыми элементами (Рис. 4б).



**Рис. 4:** Температура поверхности моря по данным со спутника Terra за 20 июля 2019 года, 08:30: (а) – исходные данные ЦКП «ИКИ - мониторинг»; (б) – результат применения алгоритма восполнения с 57456 ненулевыми элементами ( $N_{re} = 17,99\%$ )

На Рис. 5 представлена температура поверхности моря по данным со спутника Suomi-NPP за 9 апреля 2020 года, 23:24: исходные данные ЦКП «ИКИ - мониторинг» (Рис. 5а) сравниваются с результатом применения алгоритма восполнения с 22033 ненулевыми элементами разреженной матрицы (Рис. 5б).



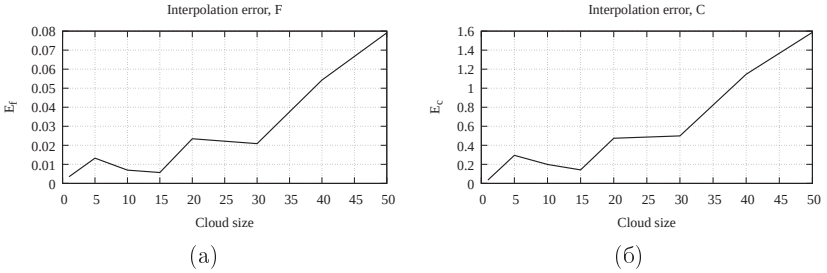
**Рис. 5:** Температура поверхности моря по данным со спутника Suomi-NPP за 9 апреля 2020 года, 23:24: (а) – исходные данные ЦКП «ИКИ - мониторинг»; (б) – результат применения алгоритма восполнения с 22033 ненулевыми элементами разреженной матрицы ( $N_{re} = 10, 84\%$ )

Для оценки качества интерполяции температуры в ненаблюдаемые зоны с использованием алгоритма восполнения были проведены следующие численные эксперименты.

В первом эксперименте исследовалась зависимость ошибки интерполяции от размера ненаблюдаемой области. Для этого были использованы данные реанализа за 1 января 2018 года, 12:00. Из исходных данных исключались подматрицы размерами  $N \times N$ , где параметр  $N$  варьировался от 5 до 50 точек. Левый нижний угол каждой исключаемой подматрицы соответствовал координатам  $41,806^\circ$  с.ш.,  $30,65^\circ$  в.д., а координаты правого верхнего изменялись от  $41,986^\circ$  с.ш.,  $30,9^\circ$  в.д. (для  $N = 5$  точек) до  $43,606^\circ$  с.ш.,  $33,15^\circ$  в.д. (для  $N = 50$  точек). Для каждой полученной матрицы применялся алгоритм восполнения, как описано выше. Результат применения алгоритма восполнения сравнивался с исходным (до исключения подматриц) набором данных. Для сравнения использовались величины  $E_c$ ,  $E_f$ ,  $MSE$ , при вычислении которых  $\Omega_h$  состоял только из индексов вырезанной (исключенной) подматрицы данных. Таким образом, мощность множества  $\Omega_h$  для вычисления ошибок интерполяции составляла  $N^2$ , то есть, равнялась количеству вырезанных точек.

На Рис. 6 представлены зависимости ошибок интерполяции  $E_f$  (Рис. 6а) и  $E_c$  (Рис. 6б) от размера «облака» (исключаемой подматрицы).

Как видно из графиков, при размерах пропуска меньше чем 30 точек ошибка интерполяции находится в пределах погрешности продукта (около  $0,5^{\circ}\text{C}$ ).



**Рис. 6:** Зависимость ошибки интерполяции зоны “облака”, (а) –  $E_f$ , (б) –  $E_c$  ( $^{\circ}\text{C}$ ), от характерного размера «облака» (количества узлов) для данных реанализа за 01.01.2018, 12:00

	$E_c$	$E_f$	$MSE$
2 подматрицы, Восполнение	0.707	$12.43 \cdot 10^{-3}$	$11.34 \cdot 10^{-2}$
2 подматрицы, Кригинг	0.425	$6.00 \cdot 10^{-3}$	$2.64 \cdot 10^{-2}$
4 подматрицы, Восполнение	0.753	$9.41 \cdot 10^{-3}$	$6.51 \cdot 10^{-2}$
4 подматрицы, Кригинг	0.433	$5.31 \cdot 10^{-3}$	$2.07 \cdot 10^{-2}$
6 подматриц, Восполнение	0.633	$8.72 \cdot 10^{-3}$	$5.61 \cdot 10^{-2}$
6 подматриц, Кригинг	0.433	$5.62 \cdot 10^{-3}$	$2.33 \cdot 10^{-2}$
8 подматриц, Восполнение	0.871	$9.33 \cdot 10^{-3}$	$6.39 \cdot 10^{-2}$
8 подматриц, Кригинг	0.433	$5.56 \cdot 10^{-3}$	$2.27 \cdot 10^{-2}$
10 подматриц, Восполнение	0.905	$9.22 \cdot 10^{-3}$	$6.22 \cdot 10^{-2}$
10 подматриц, Кригинг	0.437	$5.65 \cdot 10^{-3}$	$2.34 \cdot 10^{-2}$

**Таблица 2:** Зависимость ошибок интерполяции с помощью алгоритма восполнения и с помощью обыкновенного кригинга от количества инключенных подматриц для данных со спутника Aqua за 2 сентября 2018 года, 11:05

Во втором численном эксперименте исследовалась зависимость ошибки интерполяции от количества (частоты появления) пропусков в данных. В эксперименте были использованы данные наблюдений со спутника Aqua за 2 сентября 2018 г., 11:05. Из исходных данных исключались подматрицы размерами  $5 \times 5$ , количество подматриц варьировалось от

2 до 12. Множества индексов подматриц не пересекались. Индексы подматриц  $(i, j)$  соответствовали координатам  $(\lambda_i, \phi_j)$ , причем для каждой пары индексов каждой исключенной подматрицы выполнялось условие  $31,0^\circ \text{ в.д.} < \lambda_i < 33,0^\circ \text{ в.д.}, 42,0^\circ \text{ с.ш.} < \phi_j < 44,0^\circ \text{ с.ш.}$  Таким образом, все искусственные пропуски в данных находились в пределах определенной зоны. Для каждой полученной матрицы с пропусками применялся алгоритм восполнения. Результат применения алгоритма восполнения сравнивался с исходным (до исключения подматриц) набором данных, а также с результатом применения обыкновенного кригинга со сферической вариограммой [6]. Для сравнения использовались величины  $E_c$ ,  $E_f$ ,  $MSE$ , при вычислении которых  $\Omega_h$  состоял только из индексов вырезанных (исключенных) подматриц данных. Результаты представлены в Таблице 2. Из таблицы можно сделать вывод, что интерполяция с помощью кригинга более эффективна, чем интерполяция с помощью алгоритма восполнения.

## 5 Заключение

Результаты экспериментов показывают, что алгоритм восполнения, учитывающий наличие редких выбросов в матрице [10], позволяет эффективно сжимать ТПМ. В численных экспериментах с данными со спутников доля действительных чисел, необходимая для восстановления данных с абсолютной погрешностью, не превышающей 0,5 градуса Цельсия, составляла от 8% до 20%.

Следует отметить, что полученные факторы малоранговой аппроксимации и разреженная матрица могут быть, в свою очередь, сжаты с использованием алгоритмов кодирования битовых последовательностей без потерь. Как показали эксперименты, итоговый коэффициент сжатия  $K$  в таком случае составляет от 15:1 до 37:1. Заметим также, что разреженная компонента матричной аппроксимации содержит информацию о точках, отличающихся от малоранговой компоненты на определенную величину. Такие точки можно интерпретировать как выбросы или аномалии, если величина отклонения превышает некоторое пороговое значение. Экспериментально замечено, что обнаруженные таким образом аномалии локализуются, как правило, вблизи побережья. Возможной причиной может быть не только фактическое наличие аномалий в данных в прибрежных зонах, но и специфика формирования исходной матрицы из данных, которая может потребовать коррекции.

Алгоритм восполнения, рассмотренный в настоящей работе, также может быть использован для интерполяции данных в ненаблюдаемые зоны. Как показали эксперименты, интерполяция с помощью алгоритма восполнения менее эффективна, чем с помощью кригинга.

## Авторские декларации

### Финансирование

Т. О. Шелопут и И. А. Косолапов при поддержке Российского научного фонда (проект № 19-71-20035) провели численные эксперименты (разделы 3, 4). С. В. Петров при поддержке Российского научного фонда (проект № 21-71-10072), предложил метод (раздел 2). Авторы выражают благодарность В.П. Шутяеву за полезные замечания.

### Доступность данных и программного кода

Данные о ТПМ Черного и Азовского морей могут быть получены через информационную систему See the Sea <http://ocean.smislab.ru/static/index.sht>.

### Конфликт интересов

Отсутствует.

### Вклад авторов

Вклад авторов следующий:

Т. О. Шелопут – численные эксперименты, написание текста статьи.

С. В. Петров – разработка математических методов, написание текста статьи.

И. А. Косолапов – численные эксперименты.

### Список литературы

- [1] Copernicus Marine Environment Monitoring Service. — URL: <http://marine.copernicus.eu/>, дата посещения: 14.12.2021.
- [2] *Bradley J. N., Brislawn C. M.* Wavelet transform-vector quantization compression of supercomputer ocean models // [Proceedings] DCC93: Data Compression Conference / IEEE. — 1993. — Pp. 224–233.
- [3] *Оселедец И. В.* Вычислительные тензорные методы и их применения. — Диссертация на соискание учёной степени доктора физико-математических наук, Москва, 2012.
- [4] Six years of OSI-SAF METOP-A AVHRR sea surface temperature / A. Marsouin, P. Le Borgne, G. Legendre et al. // *Remote Sensing of Environment*. — 2015. — Vol. 159. — Pp. 288–306.
- [5] *Gandin L. S., Kagan R. L.* Construction of the system of heterogeneous data objective analysis based on the method of optimal interpolation and optimal agreement // *Meteor. Gidrol.* — 1974. — Vol. 5. — Pp. 3–10.

- [6] *Webster R., Oliver M. A.* Geostatistics for environmental scientists. — John Wiley & Sons, 2007.
- [7] Numerical model of the circulation of the Black Sea and the Sea of Azov / V. B. Zalesny, N. A. Diansky, V. V. Fomin et al. // *Russian Journal of Numerical Analysis and Mathematical Modelling*. — 2012. — Vol. 27, no. 1. — Pp. 95–112.
- [8] *Jain P., Meka R., Dhillon I.* Guaranteed rank minimization via singular value projection // *Advances in Neural Information Processing Systems*. — 2010. — Vol. 23.
- [9] *Lebedeva O. S., Osinsky A. I., Petrov S. V.* Low-rank approximation algorithms for matrix completion with random sampling // *Computational Mathematics and Mathematical Physics*. — 2021. — Vol. 61, no. 5. — Pp. 799–815.
- [10] *Petrov S., Zamarashkin N.* Matrix completion with sparse measurement errors // *Calcolo*. — 2023. — Vol. 60, no. 1. — P. 9.
- [11] Processing and assimilation of observation data for the hydrodynamics model of the Black Sea and the Sea of Azov / N. B. Zakharova, T. O. Sheloput, N. R. Lezina et al. // *Journal of Physics: Conference Series* / IOP Publishing. — Vol. 2131. — 2021. — P. 022010.
- [12] *Pisano A., Fanelli C., Cesarini C. et al.* Mediterranean Sea and Black Sea Surface Temperature NRT data. Quality Information Document [Электронный ресурс]. — 2022. — URL: <http://marine.copernicus.eu>, дата посещения: 14.12.2021.