Оценка влияния сжатия с потерей качества спутниковых данных на их измерительные свойства на примере анализа работы алгоритма LERC

Прошин А.А., Бурцев М.А., Лупян Е.А.
Институт космических исследований Российской академии наук, Москва
Современные проблемы дистанционного зондирования Земли из космоса, 11-15 ноября 2024г

В последние десятилетия наблюдается практически экспоненциальный рост объемов доступных пользователям данных ДЗЗ. Это связано как с увеличением числа действующих, в том числе открытых спутниковых систем, в особенности высокого пространственного разрешения, так и с улучшением характеристик съёмочных систем и ростом количества доступных информационных продуктов.

В 2012 году в Институте космических исследований Российской академии наук был создан центр коллективного пользования системами архивации, обработки и анализа данных спутниковых наблюдений для решения задач изучения и мониторинга окружающей среды ЦКП «ИКИ-Мониторинг (http://ckp.geosmis.ru/). Архивы центра содержат многолетние ряды данных как зарубежных, так и отечественных спутников ДЗЗ. При этом их суммарный объем на текущий момент уже превышает **7 петабайт**. Организация хранения и доступа к таким большим массивам данных очень ресурсоёмка, поэтому поиск новых решений, которые могли бы позволить сократить объемы хранения данных, является очень актуальной задачей.

Алгоритмы сжатия спутниковых изображений без потерь

Для сжатия спутниковых изображений в основном используются алгоритмы сжатия без потерь, среди которых наибольшей степенью сжатия отличается алгоритм сжатия JPEG2000. Однако, время чтения таких данных во много раз больше, чем для большинства других алгоритмов сжатия, что не позволяет использовать его для задач обеспечения интерактивного доступа к спутниковым изображениям.

Большинство остальных алгоритмов, таких как LZW, DEFLATE, LZMA, ZSTD и других, построено на базе алгоритмов Лемпеля-Зива-Уэлча и применения кодов Хаффмана. Их эффективность и быстродействие более-менее сопоставимы и зависят от природы сжимаемых данных: до 2,3 для данных Int16 и до 1,3 для данных Float32.

По совокупности достигаемых степени сжатия и минимального времени восстановления данных в формате GeoTIFF для ведения архивов данных ЦКП «ИКИ-Мониторинг» был выбран метод DEFLATE с указанием дополнительных параметров «TILED=YES PREDICTOR=2|3» (2 – для целочисленных данных и 3 – для чисел с плавающей точкой)

Compression Method	Compression Speed	Size	Read Time
Deflate	1.00	1.00	1.00
DeflateP2*	0.76	0.92	1.26
JPEG2000	0.56	0.62	8.68
LZW	2.92	1.20	1.17
PNG	0.41	0.90	1.99
LERC	3.00	0.81	0.94

Compression ratio

Algorithm			Test file		
config.ini	predictor	Z-level	*byte*	*int16*	*float32*
none	-	-	1.00	1.00	1.00
packbits	-	-	1.05	0.99	0.99
deflate	1	6	2.12	1.57	1.19
deflate_pred2	2	6	2.51	2.30	1.35
deflate_pred3	3	6	-	-	1.64
deflate_zlev9	1	9	2.12	1.57	1.19
deflate_zlev1	1	1	2.06	1.55	1.19
deflate_zlev1_pred2	2	1	2.34	2.24	1.35
deflate_zlev1_pred3	3	1	-	-	1.60
lzma	-	-	2.50	1.70	1.05
lzw	1	-	1.98	1.20	0.96
lzw_pred2	2	-	2.55	2.21	1.12
lzw_pred3	3	-	-	-	1.30
zstd	1	9	1.89	1.50	1.19
zstd_pred2	2	9	2.64	2.33	1.38
zstd_pred3	3	9	-	-	1.64
zstd_zlev15	1	15	1.89	1.50	1.19
zstd_zlev1	1	1	1.66	1.39	1.19
zstd_zlev1_pred2	2	1	2.85	2.18	1.38
zstd_zlev1_pred3	3	1	-	-	1.61

Алгоритмы сжатия изображений с потерями качества. Алгоритм LERC

Так как классические алгоритмы сжатия с потерями, такие как JPEG, не позволяют контролировать величину ошибки в пикселях, то они могут быть применяться только для данных, которые используются для визуальной оценки и не предполагают «количественной» обработки и анализа.

В 2015 году стал доступен новый алгоритм сжатия с потерями LERC – Limited Error Raster Compression, который с одной стороны позволяет существенно сократить объем растровых данных, а с другой - позволяет задать максимальную вносимую ошибку «яркости» каждого пиксела . Алгоритм основан на том, что данные разбиваются на небольшие блоки (обычно 8x8), в каждом из блоков минимальное значение пиксела берётся за основу, вычисляется разность остальных значимых пикселов с ним, полученная разность делится на 2*MaxError (где MaxError заданный уровень допустимой ошибки) и округляется. Затем вычисляется необходимое для кодирования количество бит, после чего блок сжимается без потерь. В процессе работы алгоритма для повышения степени сжатия также строится маска «данные – нет данных»

Доклад посвящен анализу особенностей применения алгоритма LERC для сжатия различных типов спутниковых данных ДЗЗ. Рассматривается влияние заданной максимальной ошибки на возможность использования сжатых изображения для задач количественного анализа.

1234.1234	1241.8741	1256.2759	1267.2950
1280.8725	1248.2917	1272.7511	1279.3802
void	1222.2943	1239.3072	void
1264.9720	1250.0852	void	void

591	979	1699	2250
2929	1300	2523	2854
void	0	851	void
2134	1390	void	void

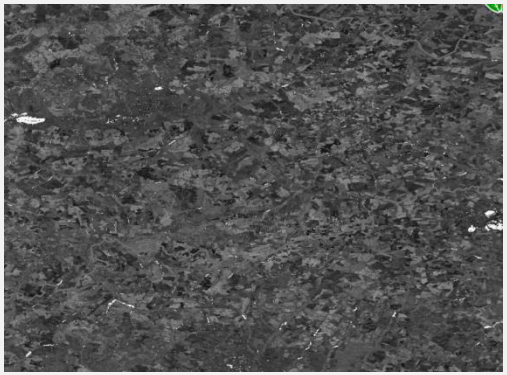
n(i) = (unsigned int)((x(i) - Min) / (2 x MaxZError) + 0.5),

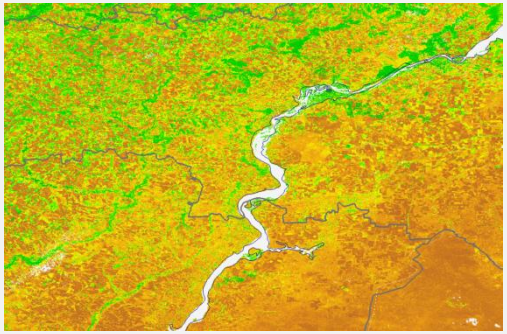
Тестовые наборы данных

В качестве основного набора тестовых данных были использованы восстановленные композитные изображения, очищенные от облачности, помех и шумов, полученные на основе данных прибора MSI, установленного на спутниках серии Sentinel 2. Эти данные уже занимают достаточно большой объем в архивах ЦКП «ИКИ-Мониторинг» (около 300Тб), а в будущем планируется увеличение количества таких продуктов во много раз. Кроме этого эти данные менее эффективно сжимаются стандартным алгоритмом DEFLATE. Данные хранятся в формате Int16, а степень их сжатия составляет всего 1,33, в то время как степень сжатия исходных сцен данных этого прибора превышает 2. Использовались фрагменты композитных изображений по каналам NIR и RED, а также композита вегетационного индекса NDVI, получаемого как разность этих каналов, деленная на их сумму.

В качестве дополнительного тестового набора использовались аналогичные информационные продукты, полученные по данным прибора MODIS (спутники TERRA, AQUA), отличающегося существенно меньшим пространственным разрешением.

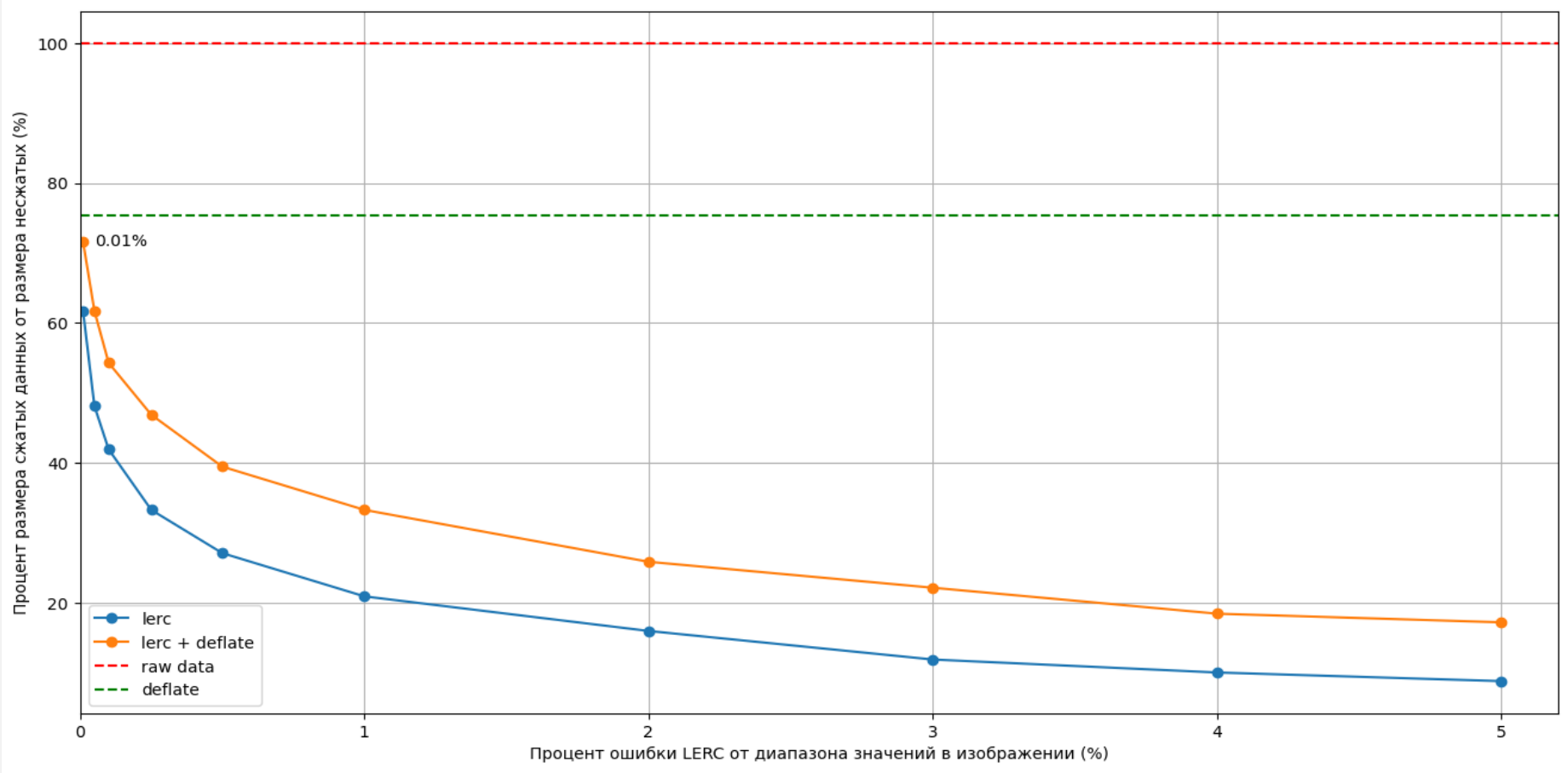
Для анализа эффективности и применимости алгоритма сжатия LERC для каждого из этих информационных продуктов была подготовлены тестовые выборки, состоящие из примерно 500 фрагментов за разные даты и с разным географическим расположением.





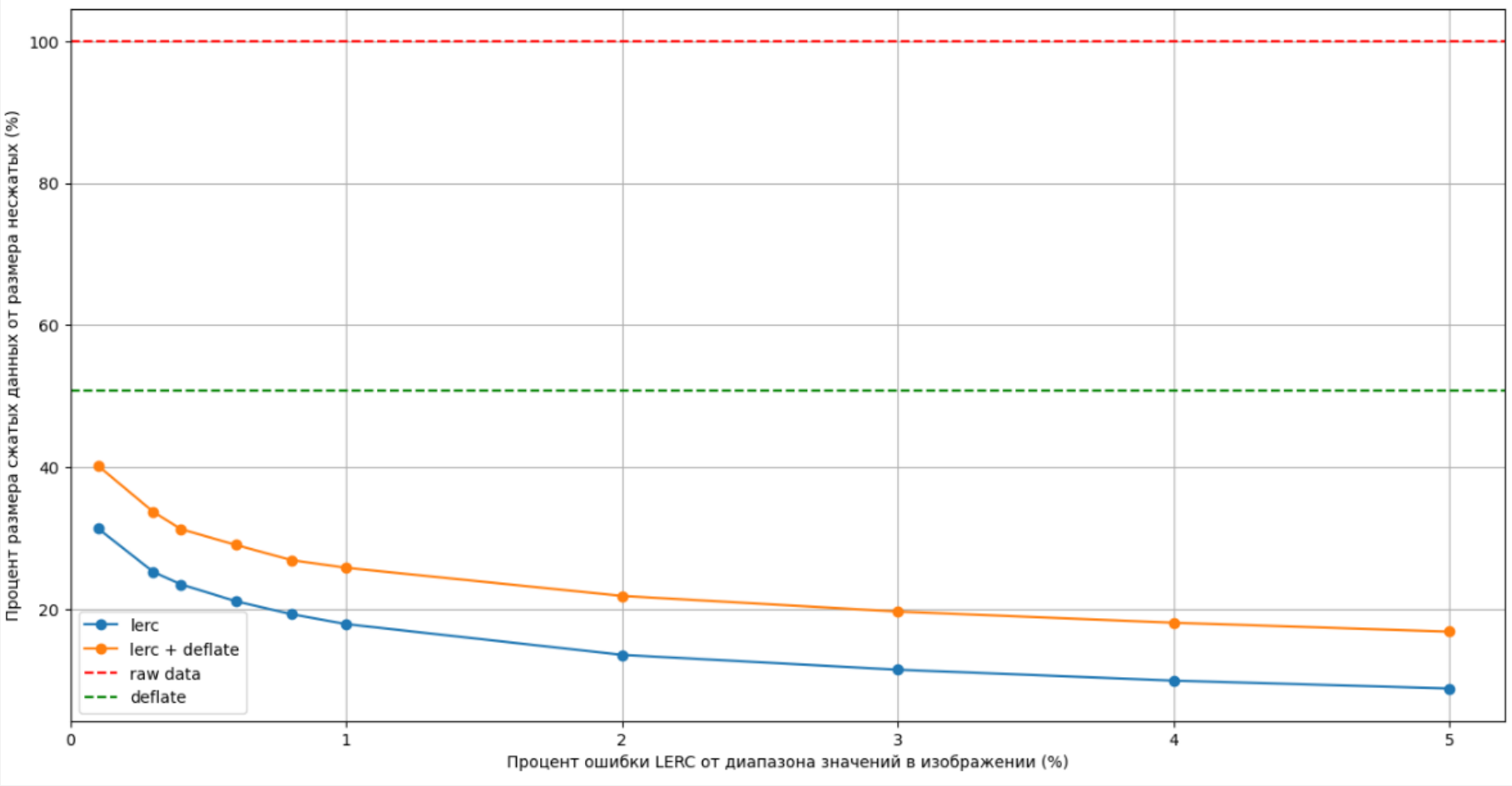
Зависимость эффективности сжатия от допустимой ошибки LERC

Для каждого фрагмента тестовых наборов были получены сжатые алгоритмом LERC изображения с использованием ряда допустимых ошибок от 0 до 5 процентов от диапазона значений в исходных изображениях. Были получены зависимости среднего объема сжатых данных от используемой ошибки LERC. Ниже приведен пример такой зависимости для восстановленных композитных изображений индекса NDVI по данным спутников серии Sentinel 2.



Зависимость эффективности сжатия от допустимой ошибки LERC

Пример зависимости для восстановленных композитных изображений индекса NDVI по данным спутников TERRA, AQUA (прибор MODIS).



Задача выбора допустимой ошибки LERC

- Для практического применения алгоритма LERC необходимо оценить порог допустимой максимальной ошибки, при которой вносимые при сжатии искажения не будут критичными для дальнейшего проведения анализа и обработки данных

- Как исходные спутниковые данные, так и продукты, получаемые на их основе, характеризуются максимальной ошибкой значений в пикселе изображения.

- Для исходных сцен данных спутников Sentinel 2 она составляет около 3%, а для используемых в выборке восстановленных композитных изображений более 5%

- Для того, чтобы при сжатии в изображения не вносилась существенная дополнительная ошибка максимальная ошибка LERC должна быть существенно меньше точности данных.

Использование спутниковых данных для получения средних по объектам мониторинга

Одним из вариантов использования спутниковых данных является получение различных интегральных характеристик по объектам мониторинга, и в частности, среднего значения по ним. Так как ошибка в каждом пикселе лимитирована и разнонаправлена, то при получении средних значений по объектам наблюдения мы ожидаемо получим гораздо меньшую ошибку. Это объясняется тем, что стандартная ошибка среднего обратно пропорциональна корню из количества элементов в выборке. При этом, чем выше пространственное разрешение спутниковых данных, тем меньше ошибка при получении средних значений по объектам наблюдения. Получается, что даже использование при сжатии LERC максимальной ошибки, к примеру, в 5% не повлияет существенным образом на вычисление средних значений по объектам мониторинга.

Стандартная ошибка выборочного среднего вычисляется по формуле:

$$\mathrm{SE}_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Использование спутниковых изображений для детектирования объектов на изображениях

Не менее важным является детектирование различных типов объектов на изображениях. В этом случае важна не столько ошибка в пикселах, сколько **сохранение пространственной структуры изображений**.

Для выявления пространственных искажений анализировались изображения, получаемые как разность между исходными данными и данными, сжатыми алгоритмом LERC с разной максимальной ошибкой. Основная гипотеза заключается в том, что если в такой разности явно прослеживается текстура, то это означает, что пространственная структура исходного изображения была искажена. Существенно, что характеристики таких изображений зависят не только от конкретных типов спутниковых данных, но и от конкретного региона наблюдения и времени наблюдения. Однако на качественном уровне все закономерности сохраняются.

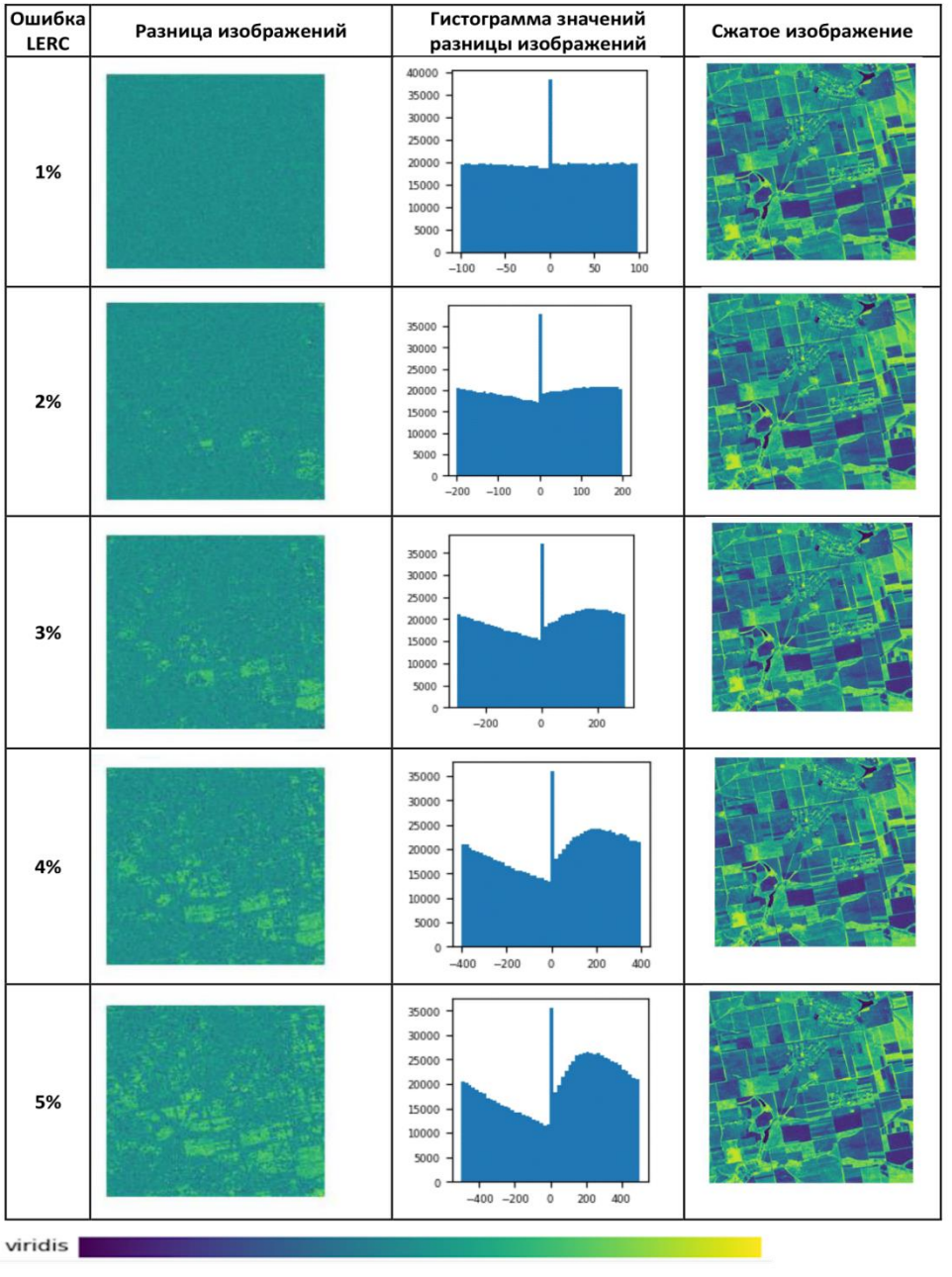
Анализ искажений пространственной структуры изображений

В качестве примера на рисунке справа рассматривается зависимость разности между оригинальным и сжатыми изображениями от допустимой ошибки для тестового фрагмента изображения 1000x1000 индекса NDVI, полученного по данным спутника Sentinel 2.

При ошибке в 1 % «разность» похожа на шум, что подтверждается гистограммой значений.

С возрастанием допустимой ошибки LERC в разнице начинает проявляться текстура, характерная для исходного изображения, а гистограмма становится несимметричной.
Это свидетельствует о том, что часть информации о пространственной структуре была потеряна при сжатии.

В то же время на самих сжатых изображениях разница хорошо заметна только при переходе на пиксельный уровень.



Основной критерий для выбора допустимой ошибки LERC

Хотя для вычисления средних значений по объектам допустимая ошибка в 3 или в 5 процентов не является существенной, она искажает пространственную текстуру, присутствующую в несжатых изображениях, поэтому **критерий сохранения пространственной структуры изображений является определяющим** при выборе максимальной ошибки при использовании алгоритма сжатия LERC.

Для практического использования рассматриваемого алгоритма сжатия необходимо выработать **количественную методику выбора допустимой максимальной ошибки LERC**, при которой вносимые текстурные искажения будут несущественными при использовании любого из рассматриваемых типов спутниковых изображений.

Оценка пространственных искажений при помощи алгоритмов классификации и кластеризации

Для выявления разных типов объектов на спутниковых изображениях (например, сельскохозяйственных полей или лесных массивов), используются различные варианты алгоритмов классификации и кластеризации. Поэтому результаты применения таких алгоритмов в первую очередь рассматривались в процессе поиска численных критериев для оценки искажений, вносимых в изображения при сжатии алгоритмом LERC. При этом, так как сжатые спутниковые изображения могут быть использованы для самых разных задач, речь шла не о выявлении каких-то конкретных объектов на изображениях, для чего обычно используются подобные методы, а о получении универсальных метрик искажения пространственной структуры.

Анализ показал, что для количественной оценки пространственных искажений **такой подход неприменим**. Результаты применения самых разных алгоритмов классификации к спутниковым изображениям без привязки к поиску конкретных объектов оказались крайне неустойчивыми и зависимыми даже от самых незначительных изменений в анализируемых изображениях.

Численные критерии для оценки искажений пространственной структуры спутниковых изображений

На основе проведенных исследований были выбраны два основных критерия для оценки искажений, вносимых в пространственную структуру изображений, рассчитываемые по разницам между исходными и сжатыми с различной ошибкой LERC изображениями:

1. Так как при повышении допустимой ошибки LERC наблюдается явная асимметрия в гистограмме значений разниц (с перекосом вправо), то в качестве первого критерия было использовано **среднее значение по разнице между исходными и сжатыми изображениями**
2. В качестве второго предположительно более точного критерия был использован **индекс автокорреляции со сдвигом 1** для ряда полученного переводом двумерной матрицы значений в одномерный массив. Индекс характеризует корреляцию между соседними элементами массива и вычисляется как коэффициент корреляции между оригинальным массивом и массивом, сдвинутым на 1 позицию. Он характеризует вероятность того, что соседние пиксели в сжатом изображении были смещены схожим образом.

Коэффициент корреляции Пирсона

Определение

Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами.

Пусть даны две выборки $x^m = (x_1, \cdots, x_m)$, $y^m = (y_1, \cdots, y_m)$; коэффициент корреляции Пирсона рассчитывается по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}},$$

где $\bar{x}, \bar{y}$ – выборочные средние x^m и y^m, s_x^2, s_y^2 – выборочные дисперсии, $r_{xy} \in [-1,1]$.

Корреляция используемых показателей (критериев)

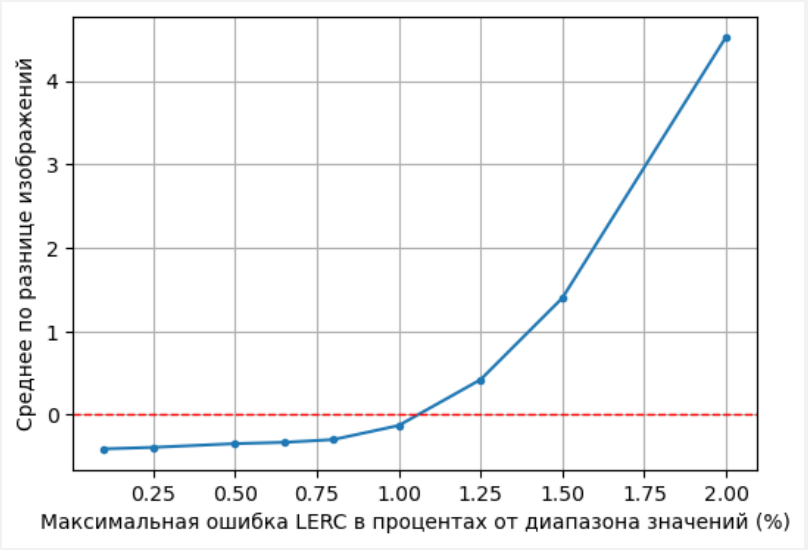
Использовался набор данных NDVI по данным спутников Sentinel 2 (диапазон значений от нуля до 10000). Для каждого фрагмента были получены: среднее по разнице и индекс автокорреляции, соответствующие различным допустимым ошибкам LERC, а также степень сжатия, вычисляемая как отношение исходного размера файла к размеру сжатого. Далее были получены агрегированные усредненные значения этих показателей для каждой ошибки LERC, для которых была построена корреляционная матрица.

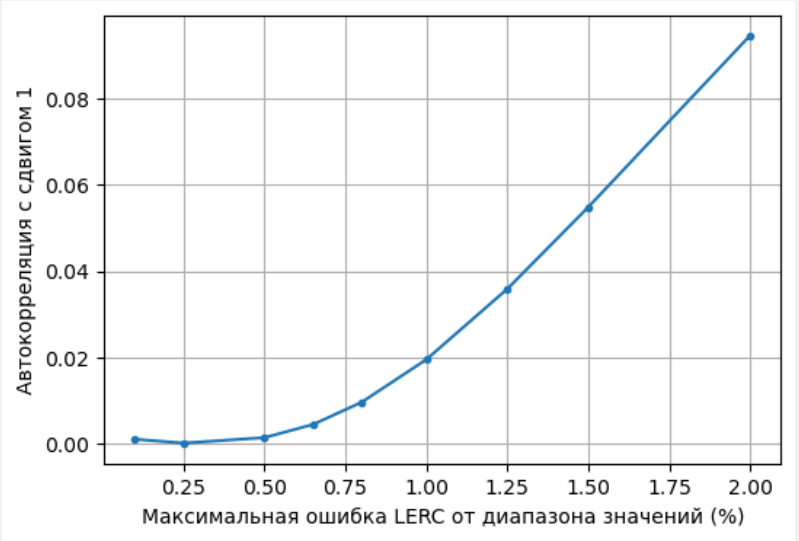
Выяснилось, что все три показателя являются почти линейно зависимыми с индексом корреляции около 0,98. Получается, что агрегированное среднее значение по разницам изображений также хорошо описывает искажения пространственной структуры изображений, как и индекс автокорреляции со сдвигом 1.

Исследование полученных зависимостей

Справа представлены зависимости усредненных значений средних по разнице и по индексу автокорреляции со сдвигом 1 для наиболее интересного для исследования диапазона ошибок LERC.

При малых значениях ошибки LERC эта величина уходит в отрицательную область. Такой эффект связан с наличием однородных областей на исходном изображении, для которых большинство значений попадает в диапазон от одной до двух ошибок от минимального значения и соответственно завышается в сжатом изображении. В то время как занижение значений в сжатом изображении при больших ошибках LERC связано с вкладом областей, для которых большинство пикселей находится в пределах одной ошибки от минимального значения.





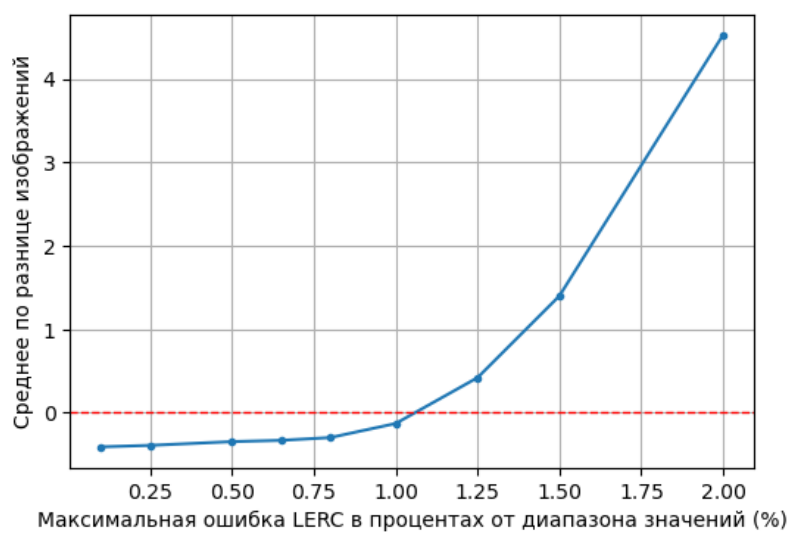
Выбор порогового значения для допустимой ошибки LERC

В качестве порогового значения для допустимой ошибки LERC было выбрано значение, при котором агрегированное среднее значение по разницам ближе всего к значению 0. Для рассматриваемого тестового набора это значение составляет примерно 1% от диапазона значений в изображениях.

Такой выбор во многом является эмпирическим и подтверждается визуальным анализом изображений сжатых с разными максимальными ошибками LERC. Этот анализ показал, что даже при просмотре изображений на пиксельном уровне при использовании указанной пороговой ошибки отличия в пространственной структуре практически не заметны.

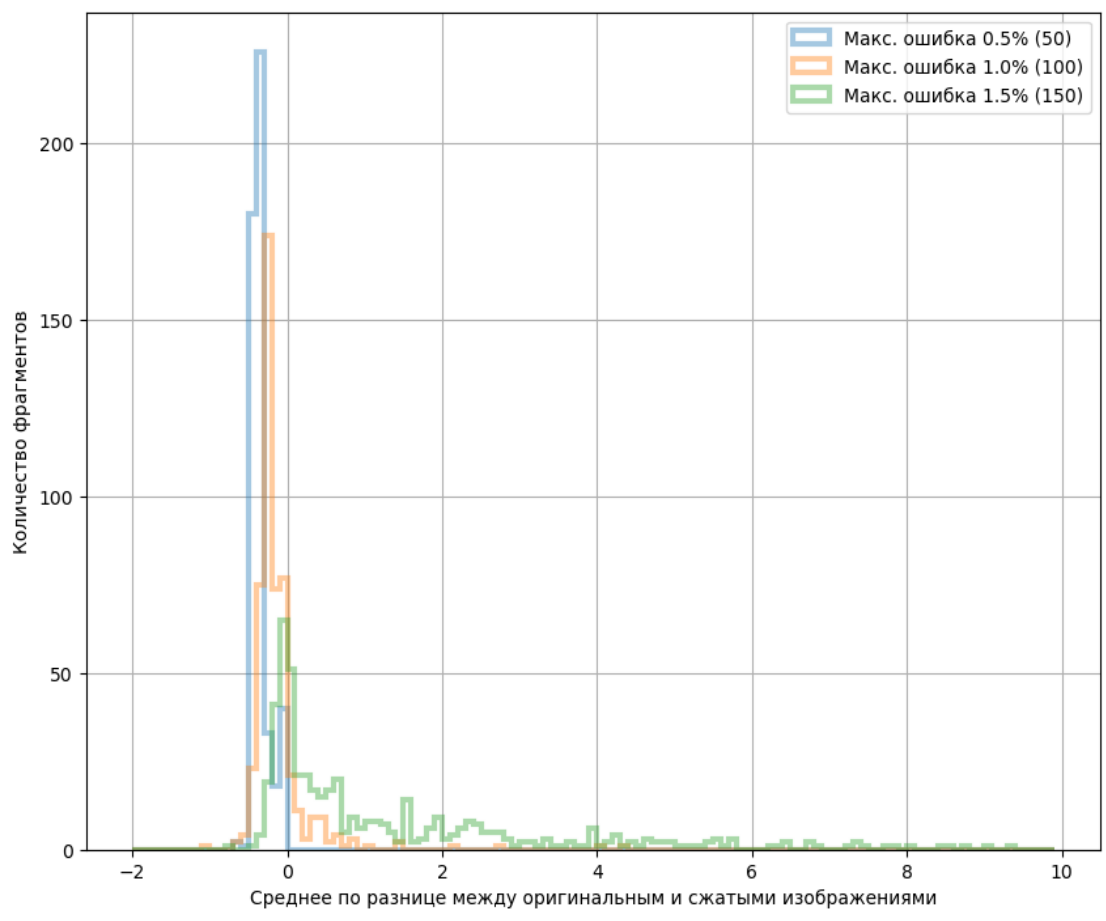
Степень сжатия алгоритмом LERC с использованием такой максимальной ошибки относительно изображений, сжатых при помощи алгоритма DEFLATE, равна примерно 3,3 (т.е. данные сжимаются более, чем в три раза).

Sentinel 2 NDVI



Анализ распределения среднего значения по разнице для файлов в выборке

В качестве дополнительного обоснования ниже приведена гистограммы средних значений по разницам для файлов в выборке для трех различных ошибок LERC. При увеличении ошибки от 1 до 1,5 процентов резко увеличивается дисперсия распределения. При еще больших значениях допустимой ошибки тенденция резкого увеличения дисперсии сохраняется. Это означает, что для значительной части изображений в тестовой выборке пространственные искажения существенно выше, чем в среднем.



Анализ достаточности используемой выборки

Также была проведена оценка достаточности выборки из 500 элементов для получения распределения значений индекса автокорреляции. Для этого была использована упрощенная формула для оценки минимального размера выборки для достижения требуемой точности при определении среднего по ней значения:

$MSS = ((2 * S)/P)^2$, где S – стандартное отклонение, а P – требуемая точность.

Из этой формуле можно получить выражение для получения точности для известного размера выборки:

$$P = \frac{2S}{\sqrt{MSS}}$$

Для распределения средних значений, соответствующих выявленному пороговому значению ошибки в 1% (абсолютное значение ошибки 100) дисперсия средних по разницам изображений равна примерно равна 0,4. Согласно приведенной формуле выборка из 500 элементов позволяет получить среднее по этой величине с точностью лучше, чем 0,036.

Для достижения же вполне приемлемой точности в 0,05 достаточно и вдвое меньшей выборки. Можно предположить, что и для других типов спутниковых изображений оценки необходимого размера выборки будут похожими.

Алгоритм действий для применения разработанной методики

Для определения порогового значения ошибки LERC предлагается выполнить следующие шаги:

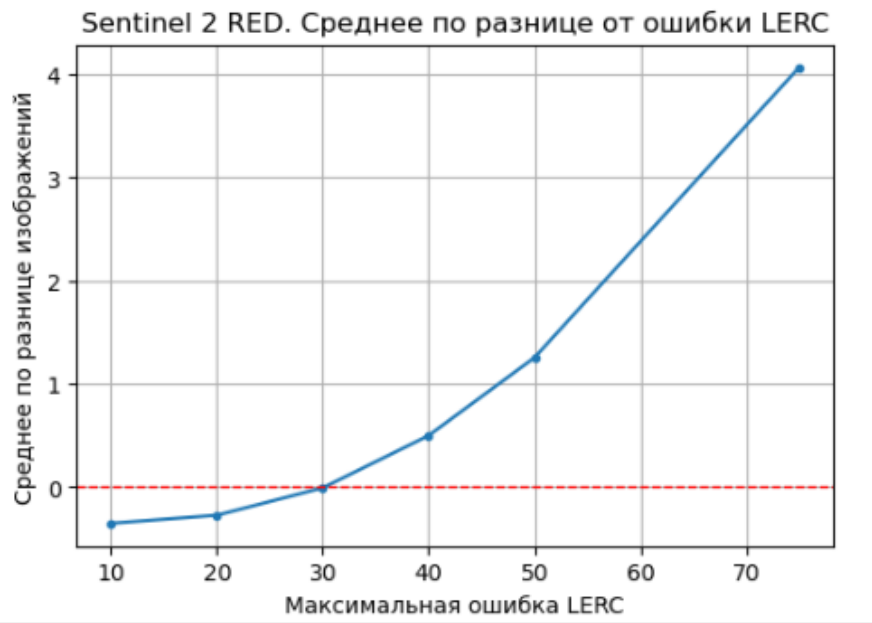
1. Подготовить случайную по времени и географическому расположению выборку спутниковых изображений анализируемого типа, состоящую из нескольких сотен экземпляров

2. Для нескольких экземпляров из этой выборки получить набор сжатых алгоритмом LERC изображений для ряда допустимых ошибок LERC в диапазоне от 0,5 до 2 процентов от диапазона значений в изображениях

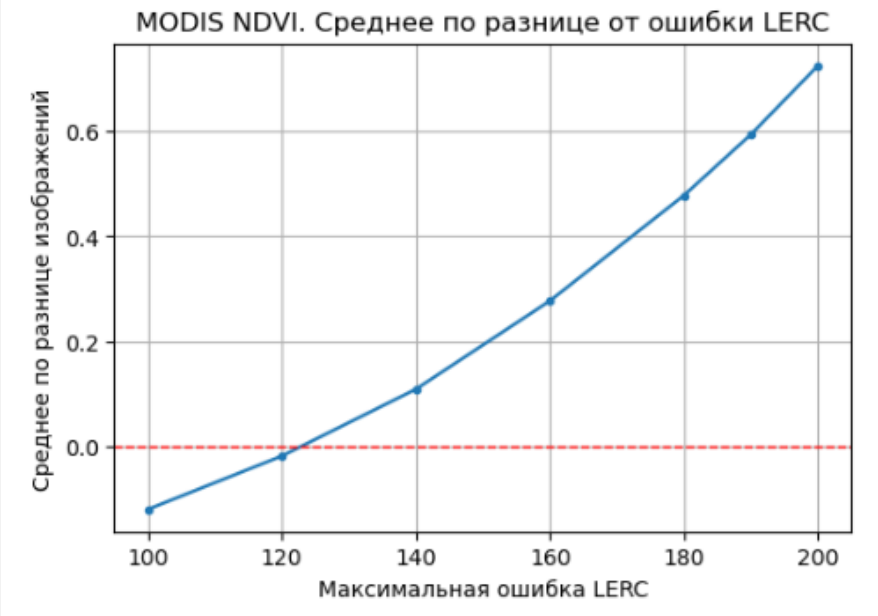
3. Получить зависимость агрегированного среднего по разницам между оригинальными и сжатыми изображениями от максимальной ошибки LERC и выбрать пороговое значение, как значение, при котором эта величина максимально близка к нулю.

4. На основании предварительного порогового значения выбрать уточненный ряд из небольшого количества различных ошибок LERC, для которых затем выполнить шаги 2 и 3 для всей подготовленной выборки и таким образом получить искомую величину

Пример использования методики для разных типов спутниковых продуктов



Допустимая ошибка LERC: 30
Относительная степень сжатия: 3,32



Допустимая ошибка LERC: 120
Относительная степень сжатия: 2,9

Заключение

- На основе анализа применимости алгоритма сжатия изображений с потерями LERC было установлено, что ключевым критерием, влияющим на выбор порогового значения является сохранение пространственной структуры изображений

- Представленная методика позволяет оценить максимально допустимую ошибку LERC для конкретного типа спутниковых изображений на основе анализа разниц между оригинальными изображениями и сжатыми с разными ошибками изображениями

- Для анализируемых типов спутниковой информации использование алгоритма сжатия LERC с указанием полученного порогового значения позволяет уменьшить занимаемый в архивах объем примерно в 3 раза

- Работы по выработке новых подходов к хранению спутниковых данных в архивах выполняются в рамках темы Минобрнауки РФ «Большие данные в космических исследованиях: астрофизика, солнечная система, геосфера» (№122042500019-6) с использованием возможностей ЦКП «ИКИ-Мониторинг» (http://ckp.geosmis.ru/)