Результаты оценки применимости сжатия с потерями LERC для данных ДЗЗ в задачах мониторинга

Прошин А.А., Бурцев М.А., Лупян Е.А., Трошко К.А., Кашницкий А.В.

Двадцать третья международная конференция
"СОВРЕМЕННЫЕ ПРОБЛЕМЫ ДИСТАНЦИОННОГО ЗОНДИРОВАНИЯ ЗЕМЛИ ИЗ КОСМОСА"
10 - 14 ноября 2025 г. в Москве

Актуальность эффективного сжатия спутниковых данных

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 9 **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)

Алгоритм сжатия изображений LERC (Limited Error Raster Compression)

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, который с одной стороны позволяет радикально сократить объем растровых данных, а с другой - 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»

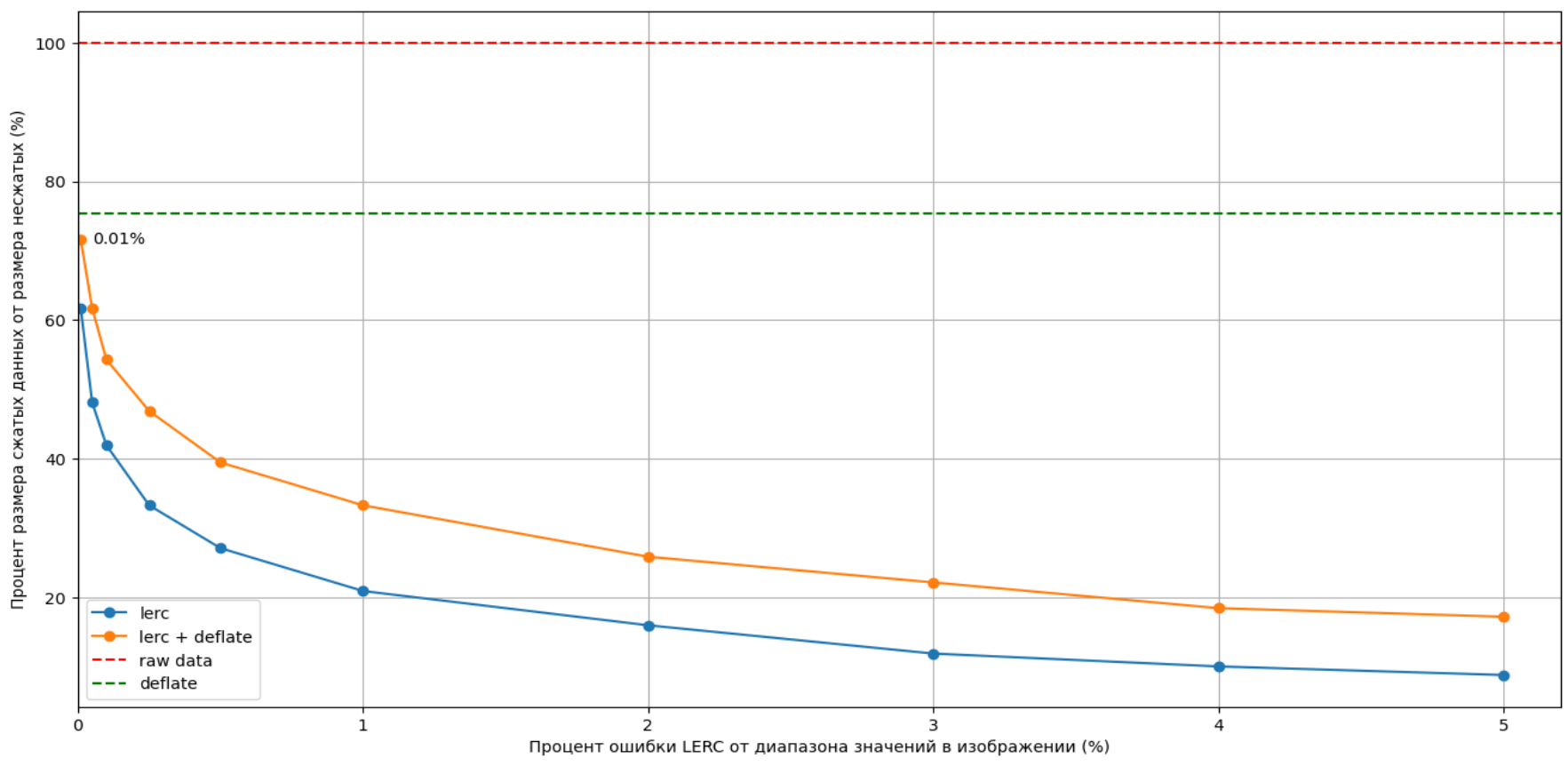
1234.1234	1241.8741	1256.2759	1267.2950
1280.8725	1248.2917	1272.7511	1279.3802
void	1222.2943	1239.3072	void
1264.9720	1250.0852	void	void

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),

Зависимость эффективности сжатия от допустимой ошибки LERC

В качестве примера ниже показана зависимость эффективности сжатия алгоритмом LERC от уровня максимальной ошибки (в процентах от диапазона значений) для восстановленных композитных изображений индекса NDVI по данным спутников серии Sentinel 2. При максимальной ошибке в 1% достигается сжатие более, чем в 3 раза. Схожие зависимости были получены и для других типов данных, включая канальные данные.



Задача выбора допустимой максимальной ошибки LERC

Оценить порог допустимой максимальной ошибки LERC можно на основе оценок радиометрической точности исходных данных. К примеру для спутников серии Sentinel-2 она составляет около 3%, а для упомянутого выше восстановленного ряда индексов NDVI – более 5%. Очевидно, что максимальная ошибка LERC должна быть существенно меньше точности данных и для точности вычисления интегральных характеристик по различным объектам мониторинга этого условия вполне достаточно.

Однако, более жестким и определяющим условием оказывается требование на **сохранение пространственной структуры изображений**, что необходимо для правильного детектирования самих объектов мониторинга.

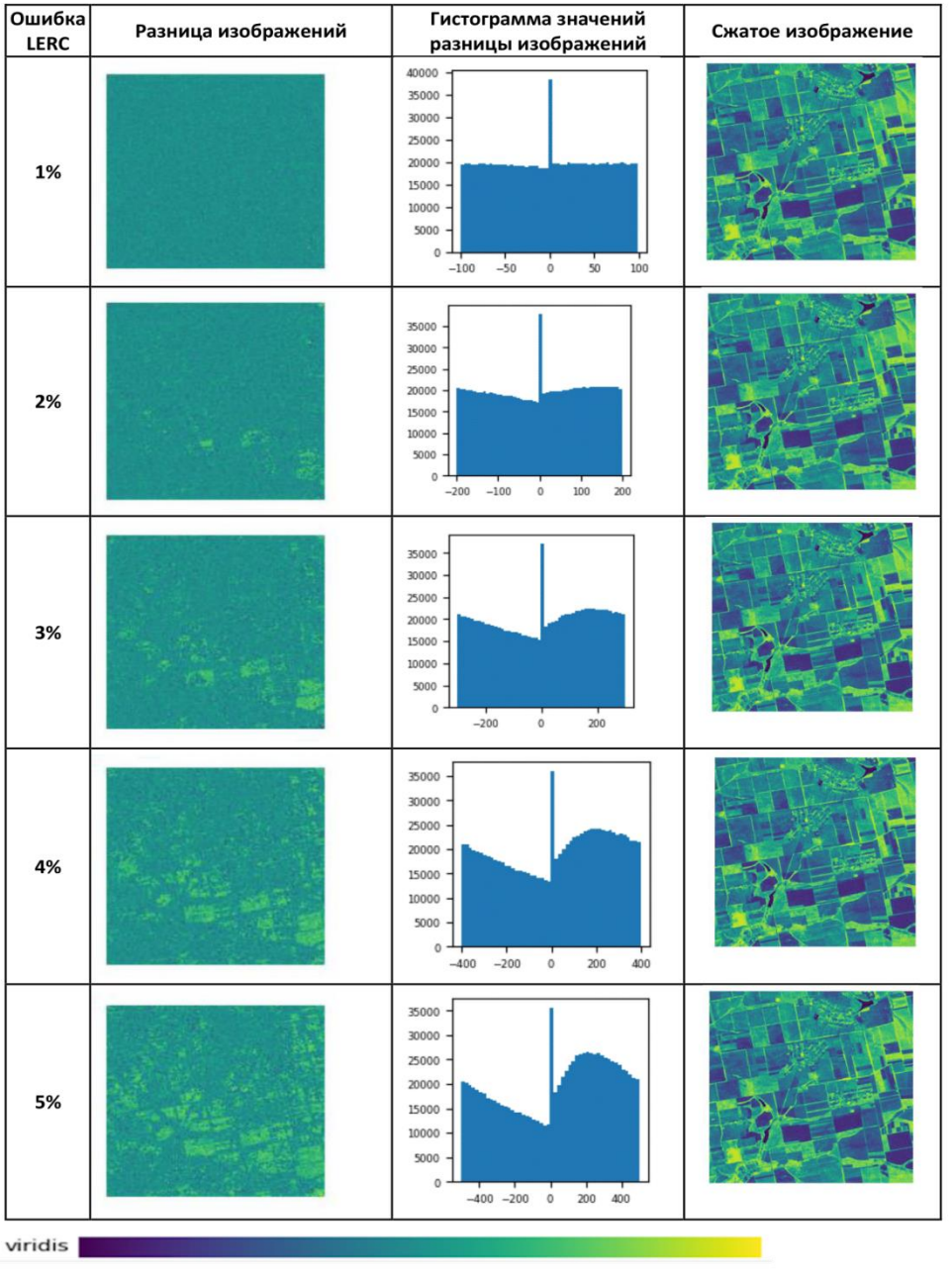
Анализ искажений пространственной структуры изображений

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.



Критерий для выбора допустимой ошибки LERC

В 2024 году в ИКИ РАН на основе проведенных исследований был выработан критерий для выбора допустимой ошибки LERC, который может быть применен к различным типам спутниковой информации.

В качестве подходящего показателя, для расчета которого не требуются большие вычислительные мощности, сначала был выбран индекс автокорреляции изображения со сдвигом 1. По сути, этот индекс характеризует вероятность того, что соседние пиксели в сжатом изображении были смещены схожим образом, что и означает наличие текстуры в анализируемой разнице. Дальнейшие эксперименты показали, что этот индекс очень коррелирует со средним значением изображений, полученных как разность между исходными и сжатыми данными, которая может быть получена с наименьшими затратами.

Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами.

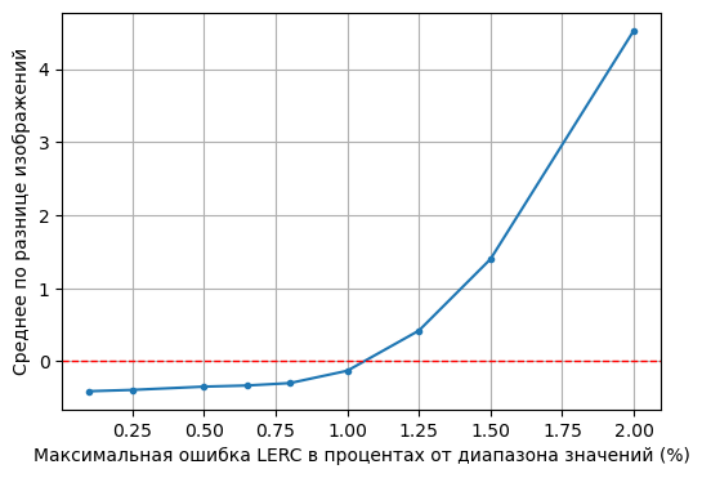
Пусть даны две выборки $x^m = (x_1, \cdots, x_m)$, $y^m = (y_1, \cdots, y_m)$; коэффициент корреляции Пирсона рассчитывается по формуле:

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

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

Критерий для выбора допустимой ошибки LERC

Ниже представлена зависимость агрегированного по тестовой выборке среднего значения по разницам между исходными и сжатыми изображениями от максимальной ошибки LERC. Видно, что при малых значениях ошибки LERC среднее уходит в отрицательную область. Такой эффект связан с наличием на исходном изображении однородных областей, для которых большинство значений попадает в диапазон от одной до двух ошибок от минимального значения в блоке и, соответственно, завышается при сжатии. 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 в блоке. В качестве порогового значения для допустимой ошибки LERC было выбрано значение, при котором среднее значение ближе всего к значению 0. В этом случае эффекты занижения и завышения значений в пикселах в результате применения алгоритма LERC максимально компенсируют друг друга.



Цель исследования

Для многих задач в настоящее время особенно востребованы восстановленные временные ряды безоблачных данных. Так как соседние по времени безоблачные наблюдения по заданной территории могут отстоять друг от друга как на дни, так и на недели, то отсутствие предсказуемой периодичности затрудняет использование исходных сцен данных для задач автоматизированного анализа и обработки данных.

В отделе «Технологии спутникового мониторинга» была разработана и используется в течении уже многих лет технология построения восстановленных ежедневных рядов данных, получаемых на основе интерполяции имеющихся безоблачных наблюдений. Ранее было показано, что сжатие таких данных при помощи алгоритма LERC с ошибкой, полученной по вышеприведенному критерию, существенным образом не влияет на качество данных.

Основной целью настоящего исследования является оценка влияние сжатия исходных канальных данных с помощью алгоритма LERC на качество производных продуктов, в первую очередь, на применимость построенного по ним восстановленного ряда данных вегетационного индекса NDVI. Выбор был связан как с востребованностью таких данных, так и с тем, что они получаются на основе формулы, которая может приводить к существенному возрастанию ошибки даже при малых ошибках в каналах. Кроме того, подобные формулы используются и для многих других индексов, получаемых по канальным данным.

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

Методика исследования

В качестве тестового набора использовались данные каналов RED и NIR прибора MSI по Рязанской области за 2024 год. На основании использования вышеописанного критерия для каждого из каналов была определена максимальная ошибка LERC и затем были получены соответствующие «сжатые» наборы данных. NIR – ошибка 60 (сжатие в 2.7 раза), RED – 40 (сжатие в 2.8 раза).

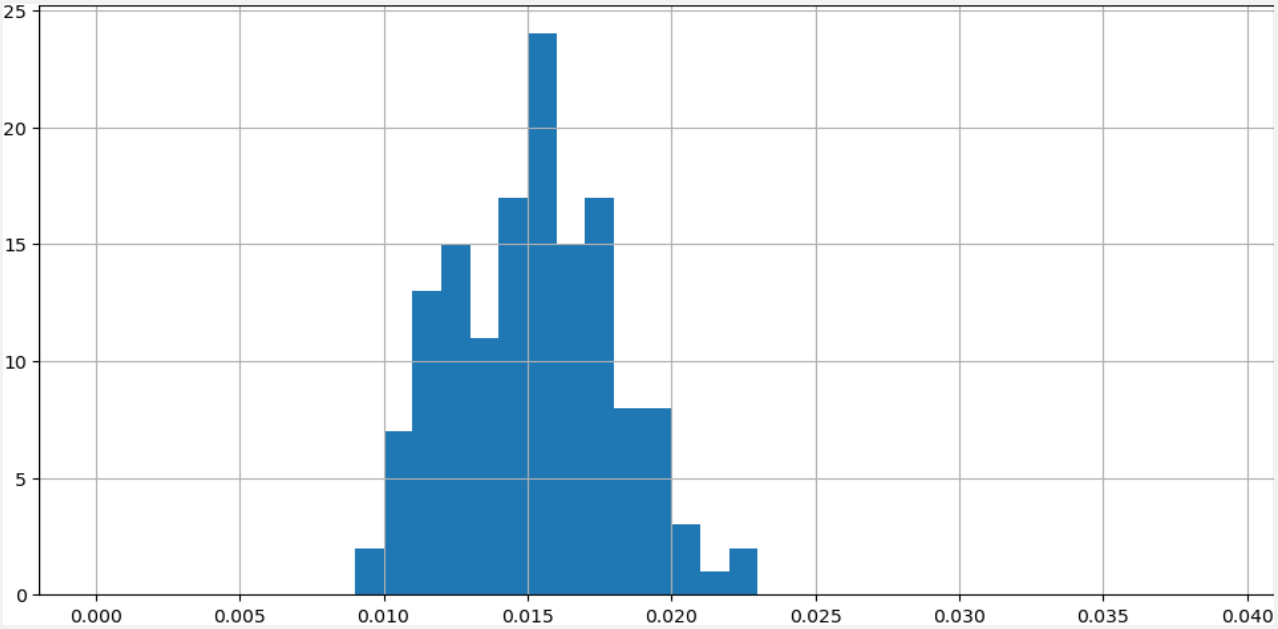
Затем на основе исходных и сжатых канальных данных были получены два соответствующих восстановленных ежедневных ряда индекса NDVI. При визуальном сравнении индексов, полученных по сжатым и несжатым канальным данным, в некоторых случаях наблюдается заметная разница в мелкомасштабной текстуре, а также в значениях отдельных пикселей. Необходимо было установить, насколько наблюдаемые отличия влияют на качество получаемого по сжатым данным ежедневного индекса NDVI или же отличия находятся в пределах его ошибки.

Для учёта влияния шумов на данных, обусловленных наличием водных объектов, снега, облаков и их теней, все сравнения проводились только по чистым участкам территории. Для этого были использованы маски классификации SCL, входящие в состав стандартного атмосферно скорректированного продукта MSI L2A и позволяющие отделить чистые области на изображениях.

Сравнение индексов NDVI, получаемых по сжатым и несжатым фрагментам исходных данных

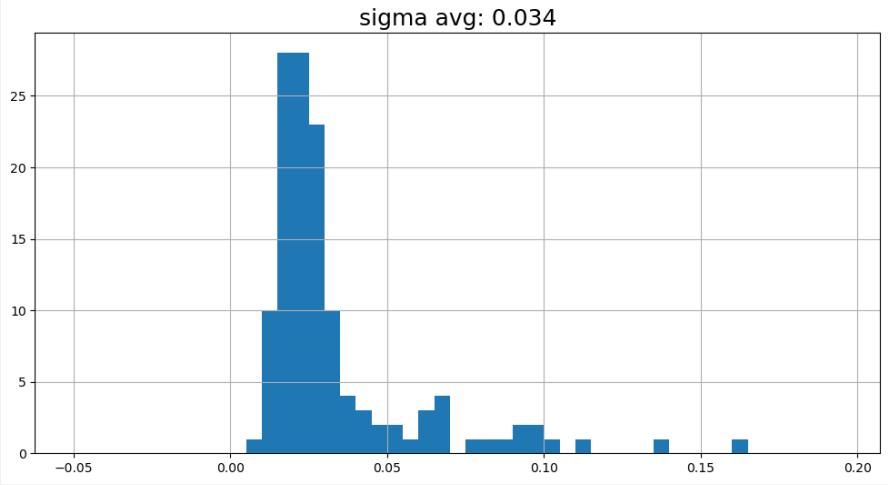
Сначала было произведено сравнение индексов NDVI, полученных по фрагментам исходных канальных данных и по сжатым при помощи алгоритма LERC. Среднее стандартное отклонение по всем сценам составило около 0.015, что составляет 1.5% от диапазона возможных значений и не превышает точности исходных данных. Ниже показана гистограмма значений стандартных отклонений, полученных по анализируемым фрагментам.

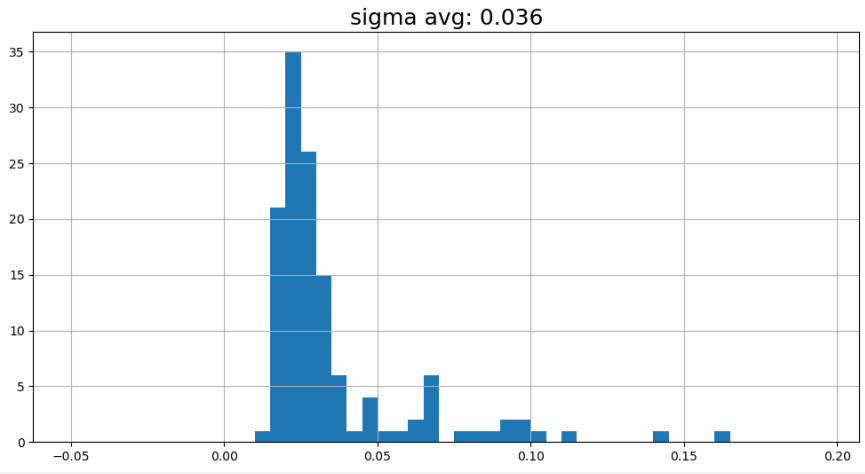
.



Сравнение индексов NDVI, полученных по фрагментам исходных данных, со значениями в восстановленных рядах данных

Изображения индекса NDVI, полученного по фрагментам исходных канальных данных были сопоставлены с соответствующими изображениями восстановленных ежедневных рядов индекса, полученных по исходным канальным данным и по данным, сжатым при помощи алгоритма LERC. В первом случае среднее по всем фрагментам стандартное отклонение составило около 0.034 (3.4%), а во втором – 0.036 (3.6%). Т.е. использование сжатых с потерями канальных данных несущественно повлияло на значения восстановленного индекса, точность которого, как уже было выше сказано не превышает 5%.





Сравнение результатов классификаций сельскохозяйственных угодий

В качестве характерной задачи, для которой применяются восстановленные ряды данных, была выбрана задача классификации сельскохозяйственных полей для определения типа их использования, а также преобладающей культуры на них.

Для проведения классификации был использован разработанный в ИКИ РАН метод, основанный на использовании классификатора Random Forest. Входными данными для него являются восстановленные ежедневные индексы NDVI с начала года до даты, на которую проводится классификация. Все имеющиеся данные за этот диапазон дат собираются в единый стек, прореживаются до каждого пятого дня, и на основе этого стека формируются векторы значений NDVI в каждом пикселе. Эти векторы и становятся основными признаками для работы классификатора.

Предварительно для проведения классификации на основе экспертного анализа спутниковых данных, доступных в системе «Вега» (цветосинтезированных изображений, в том числе многовременных, а также графиков сезонного хода NDVI), на часть района была сформирована векторная карта культур, использующаяся как для обучения, так и для оценки достоверности результатов классификации.

Сравнение результатов классификаций сельскохозяйственных угодий

Предварительно для проведения классификации на основе экспертного анализа спутниковых данных, доступных в системе «Вега» (цветосинтезированных изображений, в том числе многовременных, а также графиков сезонного хода NDVI), на часть района была сформирована векторная карта 10 культур, использующаяся как для обучения, так и для оценки достоверности результатов классификации. Для оценки достоверности результатов классификации была выполнена растеризация карты культур.

Код культуры	Культура	Площадь в эталонной карте, га	Количество полей в эталонной карте
1	Неиспользуемые земли	877,56	22
2	Земли под паром	3575,78	33
3	Озимые культуры (зерновые)	11544,39	83
4	Многолетние травы	517,49	26
5	Рапс озимый	1122,85	14
6	Рапс яровой	2084,57	14
7	Соя	5563,85	53
8	Кукуруза	2893,10	30
9	Подсолнечник	1652,13	12
10	Ранние яровые (зерновые и зернобобовые без кукурузы)	8229,68	77
ИТОГО		38061,40	38061,40

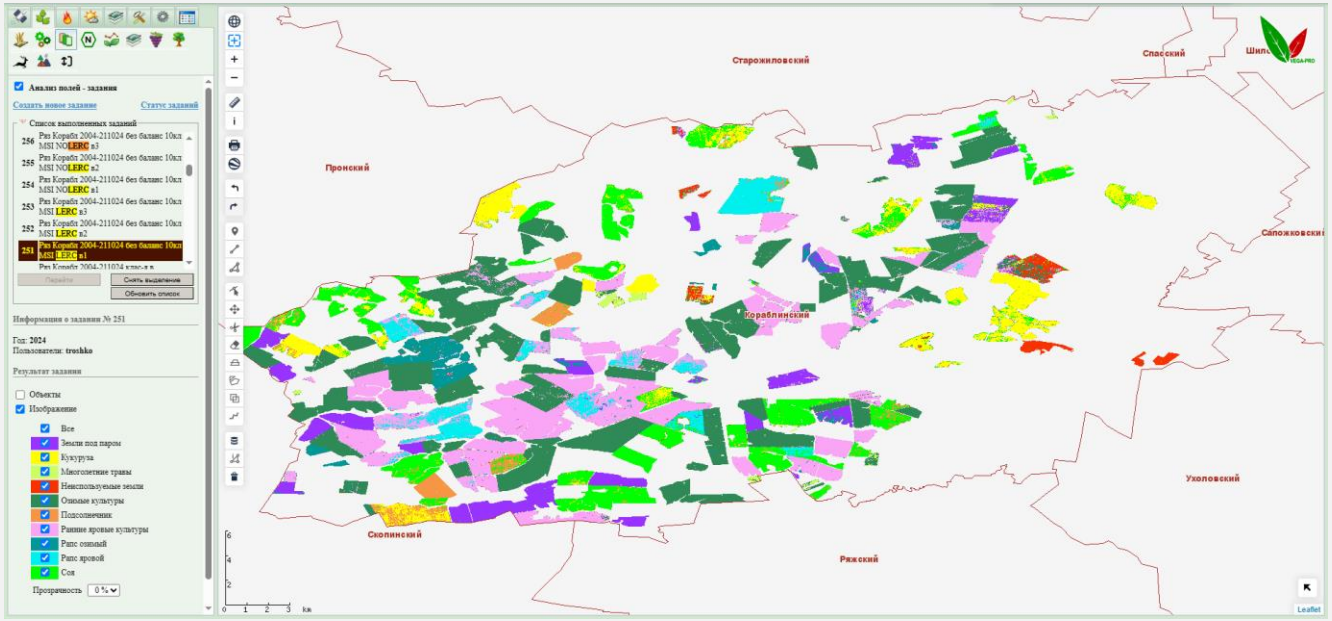
Сравнение результатов классификаций сельскохозяйственных угодий

Непосредственно попиксельная классификация изображений проводилась в пределах границ полей с использованием специализированного инструмента, реализованного в системе «Вега».

В среднем совпадение по 10 классам было достигнуто примерно в **95%** случаев.

При сравнении с эталонными данными разница в обоих случаях оказалась практически одинаковой **~86.9%**.

Основная культура на полях определялась правильно в **91%** случаев по оригинальном ряду и в **91.4%** по ряду, полученному на основе сжатых канальных данных.



Заключение

Результаты исследования показали, что сжатие канальных данных алгоритмом LERC со специально определенной максимальной ошибкой не влияет существенно на качество получаемого по ним восстановленного временного ряда индекса NDVI. Более того, исходя из особенностей проанализированного индекса, а именно, нелинейной зависимости вносимой в него ошибки от ошибки значений в каналах, можно утверждать, что влияние сжатия на восстановленные ряды канальных данных гораздо меньше. Так как большинство широко используемых индексов получаются по схожим или даже более простым формулам, то с большой долей уверенности можно утверждать, что и для них потеря качества при использовании сжатых канальных данных окажется несущественной.

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/)