Возможности оптимизации хранения временных рядов данных ДЗЗ

Прошин А.А., Бурцев М.А.
Институт космических исследований Российской академии наук, Москва
Современные проблемы дистанционного зондирования Земли из космоса, 13-17 ноября 2023г

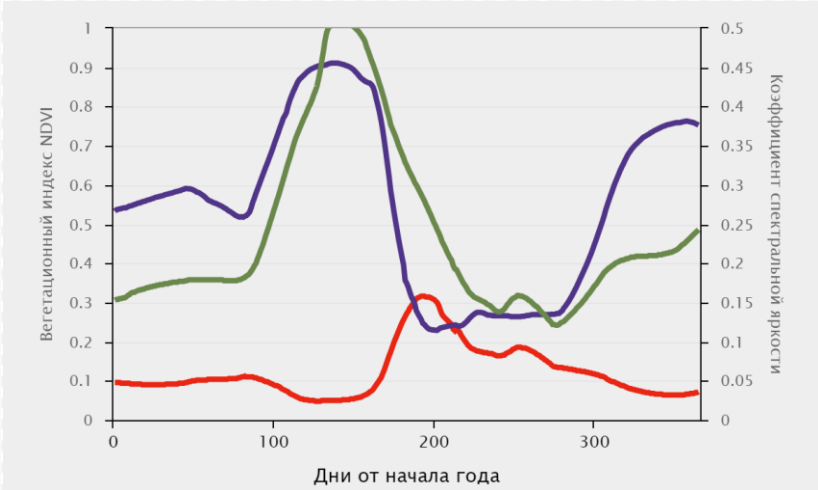
Для решения различных научных и прикладных задач в наше время все чаще используется анализ долговременных рядов информации по данным дистанционного зондирования Земли (ДЗЗ), в том числе восстановленных. В таких рядах отсутствующие по разным причинам данные, например, по причине высокой облачности, получаются на основе применения различных алгоритмов интерполяции. Характерные длины таких рядов могут составлять месяцы и даже годы, что приводит к серьёзному увеличению хранимых объемов данных. В зависимости от природы данных, объём восстановленных рядов может вырасти больше, чем в два раза относительно исходных наблюдений.

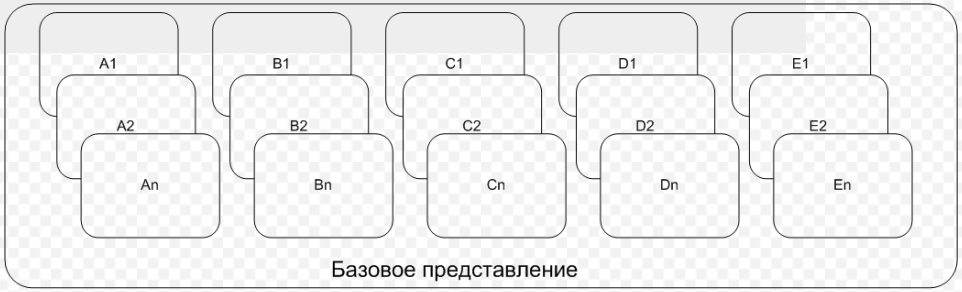
Таким образом, всё более актуальной становится задача оптимизации хранения подобных рядов данных с целью сокращения объёмов хранимой информации и, в то же время, сохранения её качества и измерительных свойств. В докладе рассматриваются два основных подхода для решения этой задачи: изменение схемы хранения данных и использование алгоритмов сжатия с лимитированной потерей качества.

«Разностная» схема хранения рядов данных

Специфика временных рядов данных такова, что они, как правило, отражают плавно изменяющиеся процессы, в ходе которых изменения каждого следующего значения относительно предыдущего сравнительно невелики.

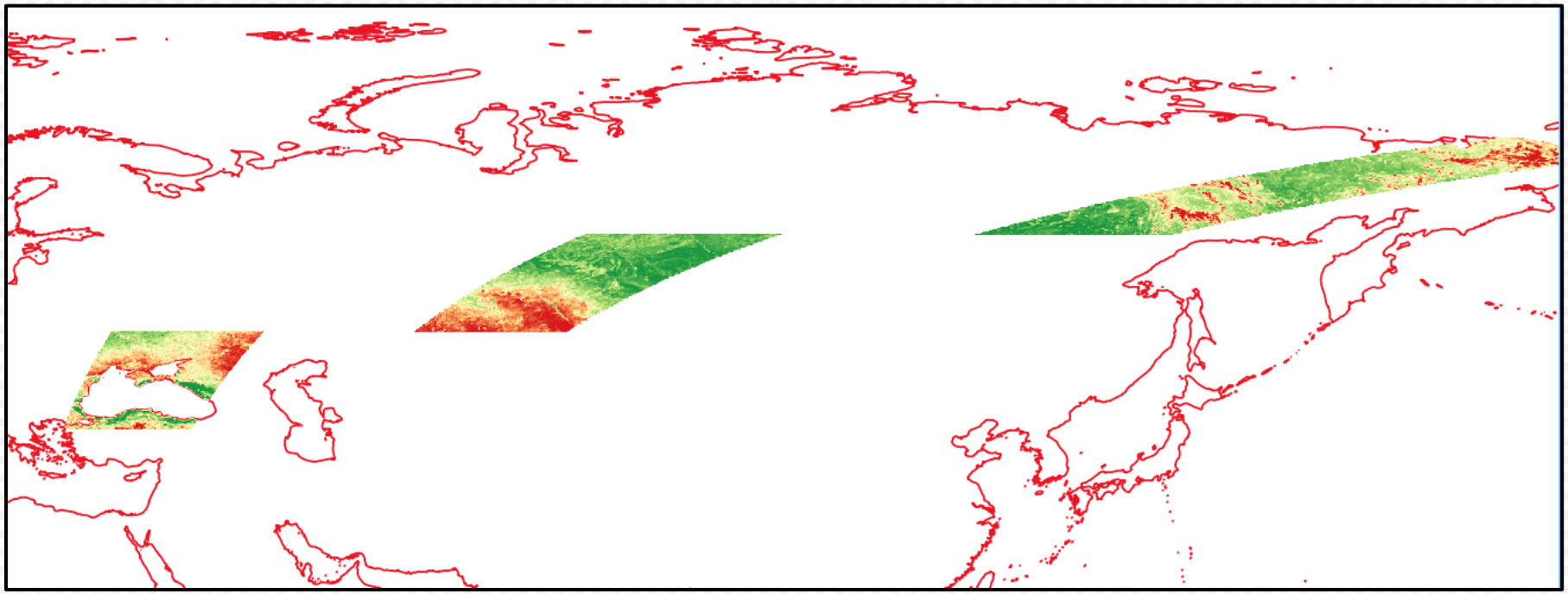
Исходя из этого эффективной оказывается т.н. «разностная» схема хранения данных, в рамках которой хранятся опорные изображения с заданным временным шагом, а в промежутках - разности между исходными и опорными изображениями. Такой подход позволяет существенно снизить объём хранимых данных за счёт резкого снижения размерности и диапазона значений у промежуточных изображений, что сильно увеличивает эффективность работы алгоритмов сжатия.





«Разностная» схема хранения рядов данных. Тестовый набор данных.

Для исследования эффективности применения «разностной» схемы хранения рядов данных в качестве тестового набора данных были выбраны ежедневные очищенные и восстановленные композиты MODIS, каналы RED, NIR, а также рассчитанный индекс NDVI. Использовались тайлы h20v04, h22v03, h24v02 за 2022 год. Выбор в первую очередь был связан с тем, что в архивах ЦКП «ИКИ-Мониторинг» содержатся временные ряды по этим данным более чем за последние 20 лет и в настоящее время они являются одними из самых востребованных при решении широкого спектра задач, связанных с мониторингом растительного покрова.



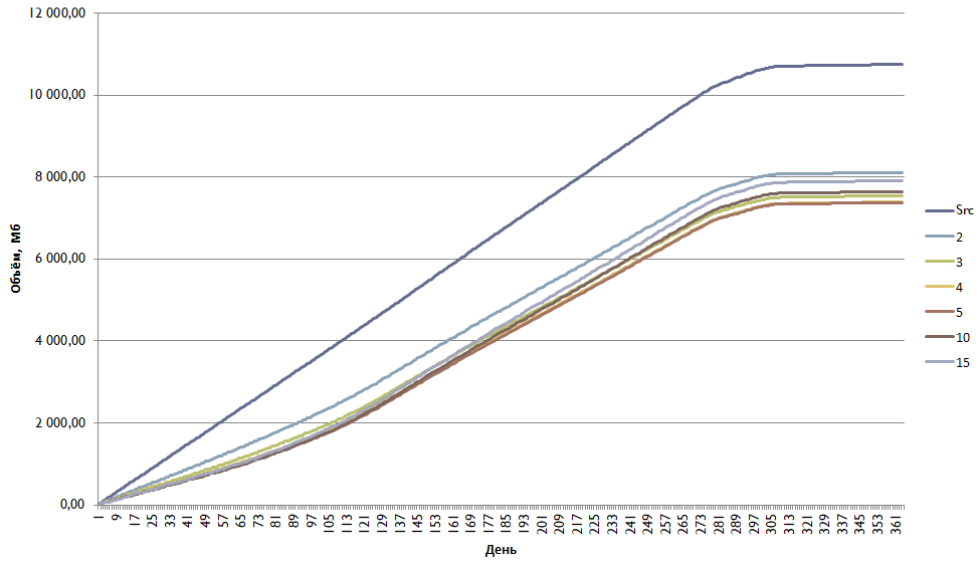
«Разностная» схема хранения рядов данных. Результаты эксперимента.

На следующих слайдах приведены результаты экспериментов по применению «разностной» схемы хранения рядов данных с использованием различного шага для опорных изображения. В таблицах приведен размер в гигабайтах и соответствующий им процент сжатия. На графиках приведен рост накопленного объема для одного из тайлов данных

Результаты - NDVI

Тайл	Исходный ряд	Шаг 2	Шаг 3	Шаг 4	Шаг 5	Шаг 10	Шаг 15
h20v04	8,92	7,25 / 81%	6,96 / 78%	6,92 / 78%	6,95 / 78%	7,23 / 81%	7,45 / 84%
h22v03	10,75	8,11 / 75%	7,54 / 70%	7,39 / 69%	7,38 / 69%	7,64 / 71%	7,91 / 74%
h24v02	8,57	6,18 / 72%	5,61 / 65%	5,42 / 63%	5,37 / 63%	5,5 / 64%	5,68 / 66%

Графики накопленного объёма для тестовых наборов – h22v03, NDVI

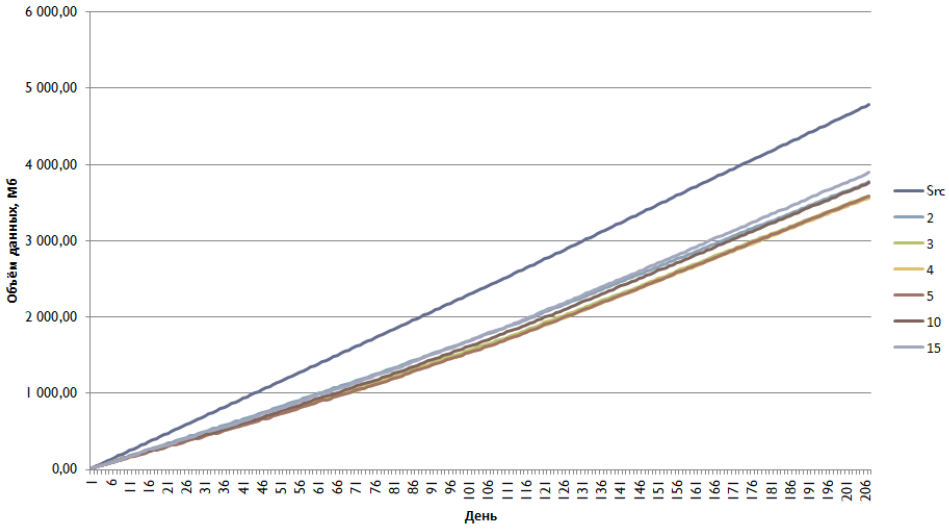


«Разностная» схема хранения рядов данных. Результаты эксперимента.

Результаты - NIR

Тайл	Исходный ряд	Шаг 2	Шаг 3	Шаг 4	Шаг 5	Шаг 10	Шаг 15
h20v04	4,79	3,78 / 79%	3,59 / 75%	3,57 / 75%	3,58 / 75%	3,76 / 78%	3,9 / 81%
h22v03	7,13	5,03 / 71%	4,54 / 64%	4,39 / 62%	4,35 / 61%	4,5 / 63%	4,68 / 66%
h24v02	6,37	4,17 / 65%	3,59 / 56%	3,38 / 53%	3,29 / 52%	3,3 / 52%	3,44 / 54%

Графики накопленного объёма для тестовых наборов – h20v04, NIR

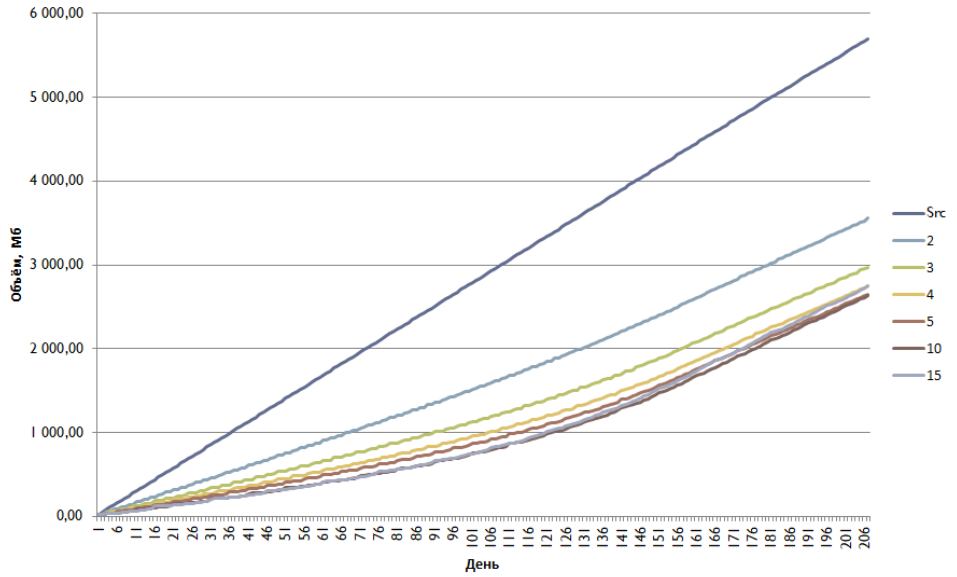


«Разностная» схема хранения рядов данных. Результаты эксперимента.

Результаты - RED

Тайл	Исходный ряд	Шаг 2	Шаг 3	Шаг 4	Шаг 5	Шаг 10	Шаг 15
h20v04	4,35	3,24 / 74%	3,03 / 70%	2,99 / 69%	3 / 69%	3,19 / 73%	3,35 / 77%
h22v03	6,53	4,37 / 67%	3,83 / 59%	3,65 / 56%	3,59 / 55%	3,71 /57%	3,89 / 60%
h24v02	5,7	3,55 / 62%	2,97 / 52%	2,75 / 48%	2,65 / 46%	2,63 / 46%	2,75 / 48%

Графики накопленного объёма для тестовых наборов – h24v02, RED



Предварительные результаты эксперимента:

- В результате применения схемы достигается экономия от 20 до 50% от исходного объёма;
- Требуется оценка быстродействия в разных сценариях – от предоставления единичных экземпляров данных до длительных процессов обработки;
- Оценка оптимального шага?

Сжатие с потерями. Алгоритм LERC

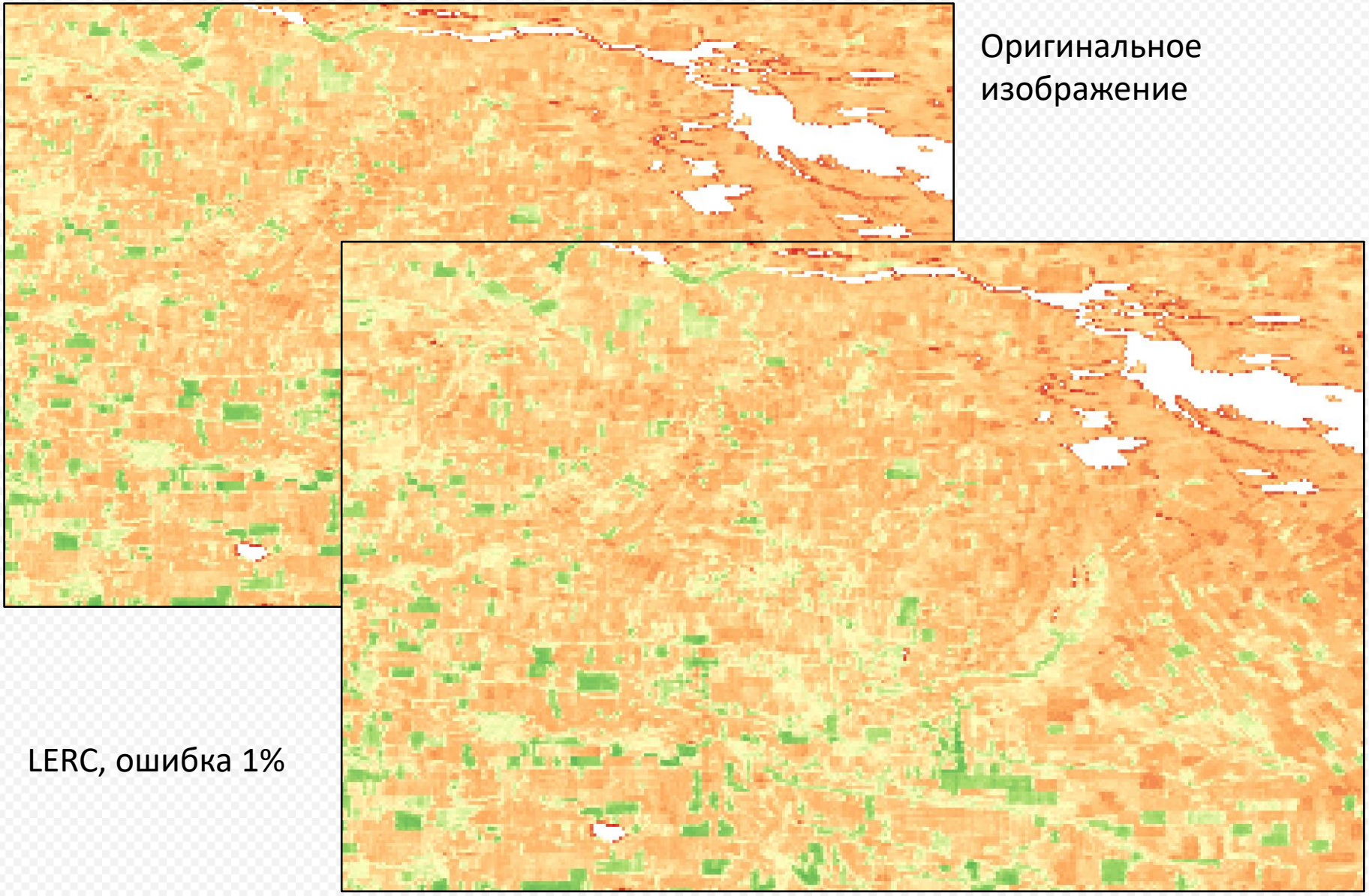
В результате анализа применяемых в настоящее время подходов для сжатия изображений с регулируемой потерей качества был выбран относительно алгоритм LERC - Limited Error Raster Compression. Ключевым преимуществом этого алгоритм является то, что он позволяет задавать целевую точность данных (MaxError), т.е. данные в каждом пикселе в результате сжатия не могут поменяться более, чем на указанную величину. В рамках применения алгоритма Данные разбиваются на блоки, в каждом блоке минимальное значение пиксела берётся за основу, вычисляется разность остальных значащих пикселов с ним, делится на 2*MaxError и округляется, после чего блок сжимается. Ниже приведена таблица с результатами эксперимента по сжатию с точностью от 1 до 5 процентов от диапазона значений.

Результаты работы LERC - NDVI

Тайл	Исходный ряд	0,01	0,02	0,03	0,04	0,05
h20v04	8,92	4,6 / 52%	4 / 45%	3,6 / 40%	3,4 / 38%	3,2 / 36%
h22v03	10,75	4,8 / 45%	4,1 / 38%	3,7 / 34%	3,5 / 33%	3,3 / 31%
h24v02	8,57	4,2 / 49%	3,6 / 42%	3,3 / 39%	3,1 / 36%	2,9 / 34%

Размеры приведены в Гб

Алгоритм LERC. Примеры изображений

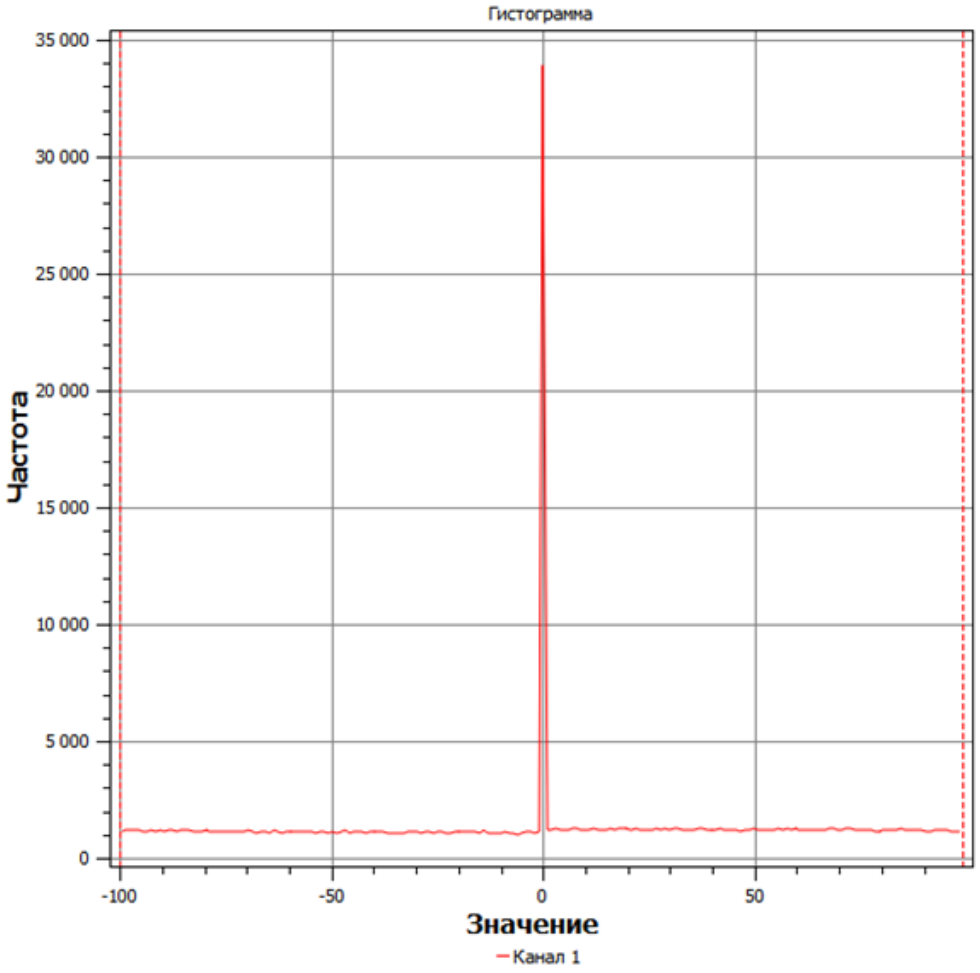


Оригинальное изображение

LERC, ошибка 1%

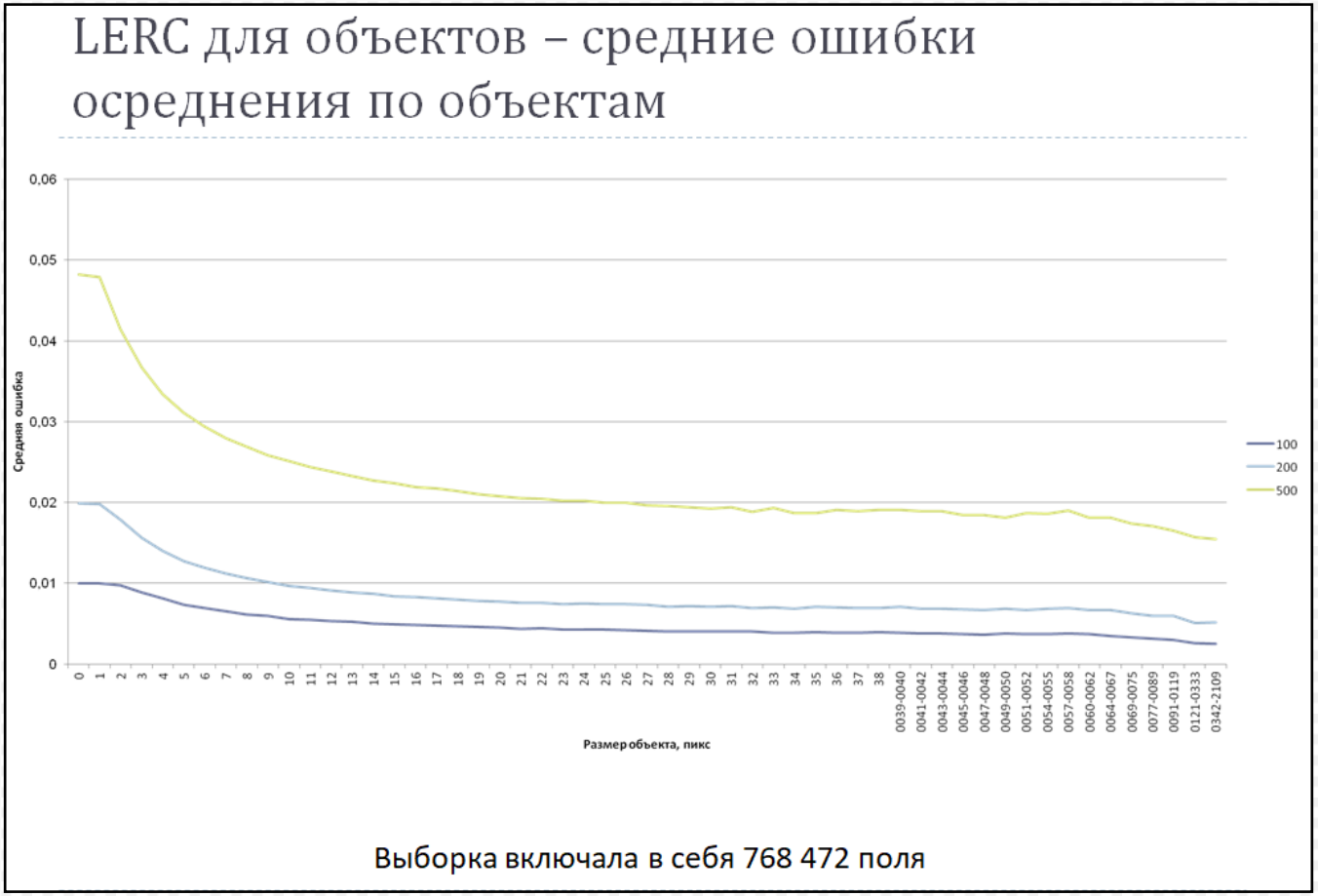
Алгоритм LERC. Гистограмма отклонений значений.

Примеры работы LERC – ошибка 0,01



Алгоритм LERC. Значения по объектам

Для многих исследовательских и прикладных задач используются интегральные характеристики по объектам наблюдения, например, по с/х полям. Ниже приводятся экспериментально полученные зависимости средней ошибки таких характеристик для объектов разного размера и разных ошибок алгоритма LERC.



Заключение

Для уменьшения объема, занимаемого в архивах долговременными рядами наблюдения, может быть эффективно применена «разностная» схема хранения данных, а также алгоритм сжатия с задаваемой максимальной ошибкой LERC. Первый подход позволяет сократить объем примерно на 20%. Применение сжатия с потерями даже с минимальными ошибками компрессии способно дать выигрыш более 50% объёма по сравнению с базовым алгоритмом DEFLATE. При этом требуется методика подбора оптимальной ошибки сжатия, при которой искажения значений в данных и текстуры изображения не повлияли бы на результаты обработки и интерпретации данных.

Работы по выработке новых подходов к хранению долговременных рядов спутниковых наблюдений 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/).