

Особенности использования алгоритма сжатия изображений LERC для архивации данных ДЗЗ

А. А. Прошин, Е. А. Лупян, М. А. Бурцев

Институт космических исследований РАН, Москва, 117997, Россия

E-mail: andry@smis.iki.rssi.ru

В последние десятилетия наблюдается практически экспоненциальный рост объёмов данных ДЗЗ. Это связано как с увеличением числа действующих, в том числе открытых, спутниковых систем, так и с улучшением характеристик съёмочных систем и ростом количества доступных информационных продуктов. Поэтому поиск новых подходов для более эффективного хранения данных ДЗЗ, в том числе за счёт совершенствования механизмов их сжатия, позволяющих сохранять требуемое качество данных и обеспечивать высокую скорость доступа к ним, включая использование алгоритмов сжатия с потерей качества, остаётся одной из актуальных задач. Настоящая статья посвящена анализу возможности использования именно такого алгоритма — LERC (*англ.* Limited Error Raster Compression), позволяющего задать максимальную ошибку в пикселе, для сжатия различных типов спутниковых данных ДЗЗ. В работе рассматривается влияние заданной максимальной ошибки на степень сжатия данных и на особенности вносимых искажений как в радиометрические свойства данных, уровень которых алгоритм позволяет контролировать, так и в пространственные характеристики сжимаемых изображений. Особое внимание уделяется анализу искажений текстуры изображений в зависимости от уровня допустимой ошибки, используемого при сжатии. Для определения максимально допустимой ошибки сжатия, не приводящей к потере не только радиометрических, но и структурных свойств изображения, предложена методика, основанная на количественной оценке изменений, наблюдаемых в текстуре изображений при различных уровнях максимальной ошибки, и приведены результаты применения этой методики для различных типов данных ДЗЗ, получаемых в видимом диапазоне, и продуктов, получаемых в результате их обработки.

Ключевые слова: спутниковые данные, ДЗЗ, архивы спутниковых данных, эффективное сжатие изображений, системы доступа к спутниковым данным

Одобрена к печати: 21.10.2024

DOI: 10.21046/2070-7401-2024-21-6-130-142

Введение

Данные дистанционного зондирования Земли (ДЗЗ) в настоящее время активно используются для решения широкого спектра научных и прикладных задач, связанных с мониторингом природной среды и антропогенного воздействия. Это стало возможным благодаря стремительному развитию спутниковых систем ДЗЗ, которое характеризуется ростом как числа различных приборов наблюдения, так и качества поступающих с них данных. При этом благодаря развитию технологий пользователям становятся доступны спутниковые данные всё более высокого и пространственного, и временного разрешения. Кроме этого для решения многих задач, связанных с наблюдением и изучением меняющихся во времени объектов и процессов, особую ценность представляют не одномоментные наблюдения, а долговременные ряды данных, что приводит к практически экспоненциальному росту объёмов архивов данных ДЗЗ.

В 2012 г. в Институте космических исследований РАН (ИКИ РАН) 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/>) (Лупян и др., 2015, 2019). Архивы Центра содержат многолетние ряды данных зарубежных и отечественных спутников ДЗЗ суммарным объёмом более 8 Пб. Организация хранения и доступа к таким большим массивам данных очень ресурсоёмка, поэтому поиск новых решений, которые могли бы позволить повысить эффектив-

ность хранения информации без потери качества результатов обработки данных, представляется очень актуальной задачей.

На сегодня реализовано большое количество алгоритмов сжатия, используемых для растровых данных (изображений) (Becker et al., 2016). Они делятся на алгоритмы с потерей и без потери качества. В областях, связанных с обработкой данных ДЗЗ, как правило применяются алгоритмы без потери качества, чтобы избежать искажения данных при расчётах. Среди таких алгоритмов наибольшей степенью сжатия отличается алгоритм JPEG2000 (<https://meetingorganizer.copernicus.org/ISRSE36/ISRSE36-593-1.pdf>) (Becker et al., 2015), но для него характерна невысокая скорость сжатия и во много раз большее время чтения, чем для большинства алгоритмов этого класса. Поэтому данный алгоритм сжатия плохо подходит для систем, обеспечивающих оперативный (особенно интерактивный) доступ к данным ДЗЗ. Большинство остальных алгоритмов, таких как алгоритмы Лемпеля – Зива – Уэлча LZW (*англ.* Lempel – Ziv – Welch), DEFLATE, LZMA (*англ.* Lempel – Ziv – Markov chain Algorithm), ZSTD (*англ.* ZSTandarD) и другие, построено на базе алгоритмов LZW и кодов Хаффмана (*англ.* Huffman coding) (Oswal et al., 2016). На текущем этапе развития их эффективность и быстродействие более-менее сопоставимы и зависят от природы сжимаемых данных. Практически все подобные алгоритмы имеют средний предел эффективности степени сжатия порядка 2,3 для данных разрядности Int16 и порядка 1,3 для данных разрядности Float32, выше которого можно перейти только в редких случаях (см., например, <https://kokoalberti.com/articles/geotiff-compression-optimization-guide/>).

Использование классических алгоритмов сжатия с потерей качества, таких как JPEG, позволяет существенно увеличить достижимую степень сжатия данных по сравнению с предыдущими алгоритмами, однако область их применения ограничивается данными, которые используются только для визуализации и не предполагают «количественной» обработки и анализа. Это связано с тем, что ошибка, вносимая в данные алгоритмом JPEG и подобными, непредсказуема и может в конечном итоге привести к существенным искажениям как данных, так и полученных результатов их обработки.

В 2015 г. стал доступен новый алгоритм сжатия с потерями — LERC (*англ.* Limited Error Raster Compression, <https://esri.github.io/lerc/>) (Becker et al., 2016), который позволяет задать и гарантировать максимальную вносимую ошибку в значение каждого пикселя (MaxError), тем самым сохраняя радиометрические свойства изображения с определённой погрешностью. Алгоритм основан на том, что данные разбиваются на небольшие блоки, в каждом из которых минимальное значение пикселя берётся за основу, вычисляется разность остальных значимых пикселей с ним, полученная разность делится на $2 \times \text{MaxError}$ (где MaxError — заданный уровень допустимой ошибки) и округляется до целого, после чего полученный блок сжимается без потерь. Таким образом достигается уменьшение размерности данных за счёт их перевода в целочисленные значения. Большая часть значений при этом изменяется, но не более чем на величину установленного порога допустимой ошибки. Результатом такого изменения является уменьшение словарей для последующего сжатия без потерь. При этом естественно, что повышение порога допустимой ошибки приводит к более существенному сокращению словарей и, как следствие, к более высоким степеням сжатия. Опыт использования данного алгоритма показывает, что даже при относительно низких значениях допустимой ошибки достигаются значительные уровни сжатия (<https://kokoalberti.com/articles/geotiff-compression-optimization-guide/>).

Настоящая статья посвящена анализу особенностей применения алгоритма LERC для сжатия различных типов спутниковых данных ДЗЗ. В ней рассматривается влияние заданной максимальной ошибки как на степень сжатия данных, так и на изменение радиометрических и пространственных характеристик сжатых данных. Показано, что увеличение вносимой ошибки может приводить к искажению пространственной структуры изображения даже при сохранении его радиометрических свойств. Для оценки подобных искажений и определения предельных значений вносимой ошибки предложена методика количественной оценки изменений, наблюдаемых в текстуре изображений при различных уровнях максимальной ошибки.

Анализ эффективности и применимости алгоритма LERC для сжатия данных ДЗЗ

Для проведения оценки эффективности и применимости алгоритма LERC для сжатия данных ДЗЗ были подготовлены наборы тестовых данных, оценены их объёмы, а также радиометрическая точность для получения максимально допустимой ошибки, после чего эти наборы были последовательно сжаты с различными степенями ошибки.

В качестве основного набора тестовых данных были использованы восстановленные композитные изображения, очищенные от облачности, помех и шумов (Миклашевич и др., 2019), полученные на основе данных прибора MSI (*англ.* MultiSpectral Instrument), установленного на спутниках серии Sentinel-2. Такой выбор был обусловлен тем, что эти данные активно используются в работах, идущих в ЦКП «ИКИ-Мониторинг», уже занимают достаточно большой объём в архивах, а в будущем планируется многократное увеличение количества таких продуктов. Кроме того, эти данные менее эффективно сжимаются стандартным алгоритмом DEFLATE. Данные хранятся в формате Int16, средняя степень сжатия для них составляет всего 1,33, в то время как степень сжатия исходных сцен данных этого прибора в среднем превышает 2, что, вероятно, объясняется наличием в исходных данных однородных участков, покрытых облачностью. Следует также отметить, что проведение анализа именно на предлагаемом типе данных позволяет оценить возможности использования алгоритма LERC при сжатии данных, свободных от облачности, которые в основном и используются при изучении различных процессов и явлений, происходящих на поверхности Земли.

Для построения выборки применялись фрагменты композитных изображений по каналам NIR (*англ.* Near InfraRed, ближний инфракрасный) и RED (красный), а также композита вегетационного индекса NDVI (*англ.* Normalized Difference Vegetation Index). В качестве дополнительного тестового набора использовались аналогичные информационные продукты, полученные по данным прибора MODIS (*англ.* Moderate Resolution Imaging Spectroradiometer) (спутники Terra, Aqua), отличающегося существенно меньшим пространственным разрешением и намного более широкой полосой обзора.

Для каждого из этих информационных продуктов были подготовлены выборки, состоящие из примерно 500 фрагментов за различные даты и с разным географическим расположением. Затем для каждого из набора данных получены сжатые алгоритмом LERC изображения с максимальной ошибкой от 0 до 5 % от исходного диапазона значений. После этого были установлены зависимости среднего объёма сжатых данных от используемой ошибки LERC. Пример такой зависимости для восстановленных композитных изображений индекса NDVI (диапазон значений от 0 до 10 000), полученных по данным прибора MSI, приведён на *рис. 1* (см. с. 133). На графике красной штриховой линией показан объём несжатых данных, принятый за 100 %, зелёной — объём данных, сжатых при помощи алгоритма DEFLATE, а синей кривой — доля (в %) от исходного объёма, получающаяся при сжатии алгоритмом LERC при разных значениях максимальной ошибки. По оси абсцисс указана ошибка как доля (в %) от возможного диапазона значений в исходных файлах. Также на графике оранжевым цветом отображена кривая, показывающая объём, получающийся после перепаковки алгоритмом DEFLATE данных, которые были сжаты алгоритмом LERC. Интерес к этой зависимости вызван тем, что алгоритм сжатия LERC поддерживается только в актуальных версиях программных пакетов работы с изображениями (прежде всего это пакет GDAL (*англ.* Geospatial Data Abstraction Library, <https://gdal.org/en/latest/>)), что может привести к проблемам с совместимостью для более старых версий программного обеспечения работы с данными. Из графика видно, что при использовании алгоритма LERC с максимальной ошибкой в 1 % от диапазона значений достигается экономия дискового пространства более чем в три раза относительно данных, сжатых алгоритмом DEFLATE, и даже при использовании компромиссного варианта — более чем в два раза относительно них же.

Аналогичная зависимость для данных композитов индекса NDVI, полученных по данным прибора MODIS, приведена на *рис. 2* (см. с. 133). Степень сжатия алгоритмом DEFLATE у этих данных оказалась существенно лучше, в то время как эффективность сжатия при

помощи алгоритма LERC отличается только для малых значений допустимой ошибки LERC. Эти композиты получаются по аналогичной методике, а отличия, судя по всему, связаны с существенно более низким пространственным разрешением этих данных.

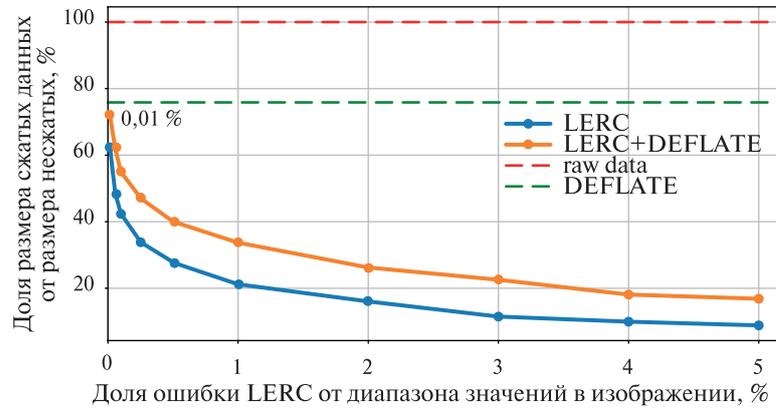


Рис. 1. Эффективность сжатия ежедневных композитов индекса NDVI, полученного по данным прибора MSI

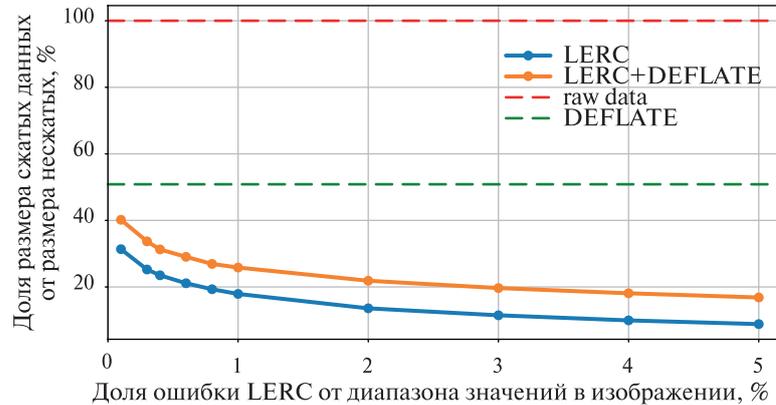


Рис. 2. Эффективность сжатия ежедневных композитов индекса NDVI, полученного по данным прибора MODIS

На приведённых графиках хорошо видно, что даже относительно небольшие значения вносимой ошибки позволяют получить степени сжатия порядка 4–5. Тем не менее для практического применения алгоритма LERC необходимо оценить порог допустимой максимальной ошибки, при которой вносимые при сжатии искажения не будут критичными при обработке данных. С точки зрения радиометрических характеристик изображения, такую ошибку можно получить на основе оценки радиометрической точности исходных данных, задав её значение, не превышающее половины исходной точности данных. Так, например, для исходных сцен данных спутников Sentinel-2 она составляет 3–5 % (см. ежегодный отчёт о функционировании системы Sentinel-2 за 2023 г., <https://sentiwiki.copernicus.eu/attachments/1673423/OMPC.CS.APR.003%20-%20S2%20MSI%20Annual%20Performance%20Report%202023%20-%20201.0.pdf?inst-v=5d892f0e-0a4f-4a11-88a9-f6f3caa8d5c0>), а для используемых в выборке восстановленных композитных изображений — порядка 5 % (Миклашевич и др., 2019). Таким образом, максимальная ошибка сжатия для этих данных будет составлять 1,5–2,5 и 2,5 % соответственно.

Вторым важным фактором при оценке допустимой ошибки сжатия является сохранение структуры изображения, так как при высоких степенях потери в силу особенностей работы алгоритма с блоками данных она заведомо портится, что влияет на решение таких задач, как детектирование объектов на изображении, расчёт средних по объектам и тому подобное.

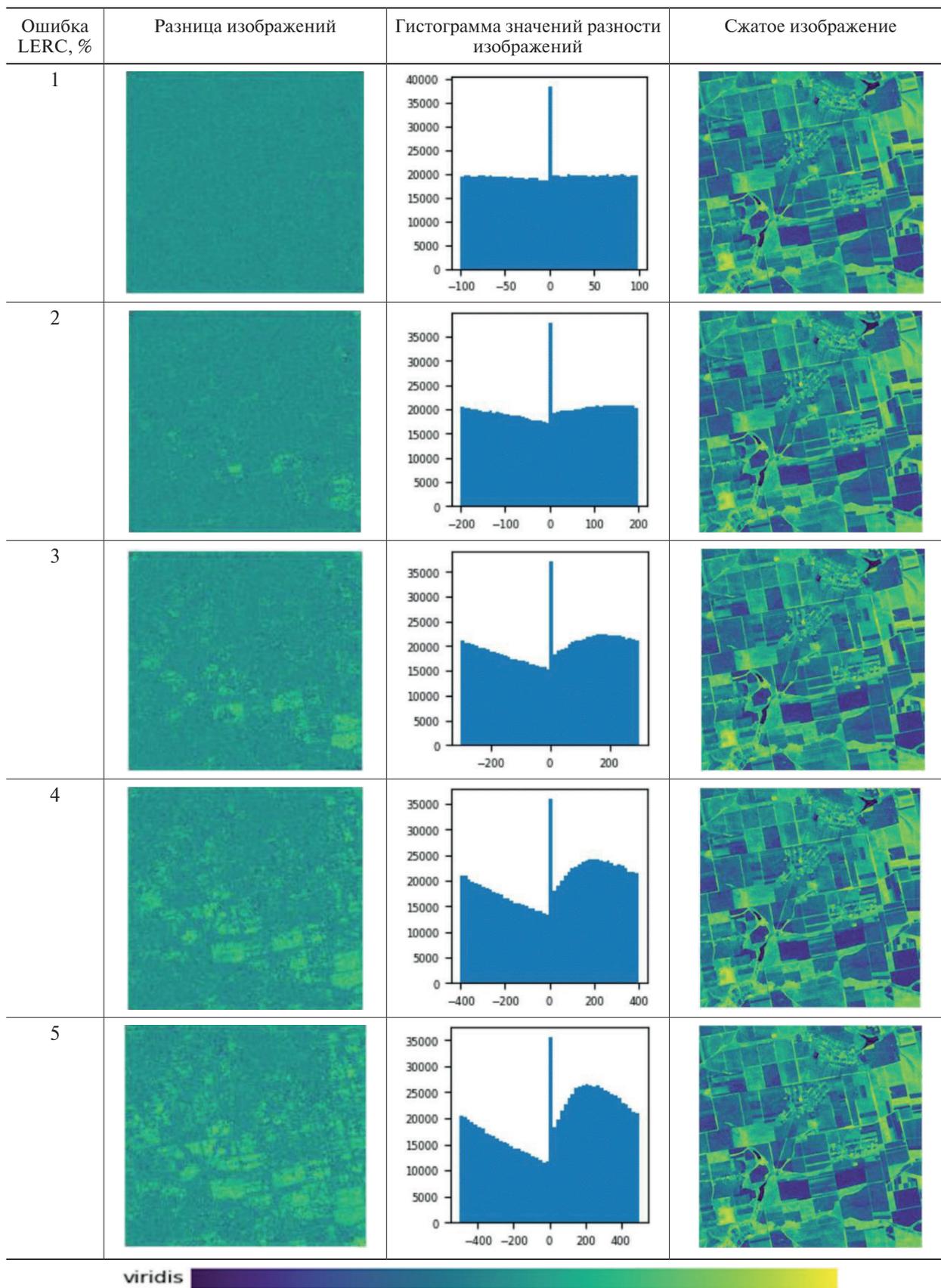


Рис. 3. Разность между оригинальными и сжатыми изображениями при использовании разных допустимых ошибок LERC и соответствующие им гистограммы значений. Для изображений используется палитра, которая пропорционально распределяется от минимальных до максимальных значений, присутствующих в пикселях изображений

Для выявления пространственных искажений анализировались изображения, получаемые как разность между исходными и сжатыми данными. Если разность близка к шуму, искажений в пространственной структуре сжатого изображения нет. Наличие в разности явной структуры является признаком того, что пространственная структура 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. Для иллюстрации изменения структуры изображения при разных ошибках сжатия ниже анализируется фрагмент изображения индекса NDVI размером 1000×1000 пикселей, полученного по данным спутников Sentinel-2.

Установленная зависимость разности между оригинальными и сжатыми изображениями от допустимой ошибки показана на *рис. 3* (см. с. 134). В первой колонке таблицы указана заданная ошибка LERC (как доля (в %) от диапазона значений), во второй — полученная разность, в третьей — гистограмма значений в ней, а в четвёртой — соответствующее сжатое изображение. При ошибке в 1 % изображение, соответствующее разности, близко к шуму, что подтверждается его гистограммой. Начиная с ошибки в 2 %, на разности начинают проявляться однородные области (структура), а гистограмма перестаёт быть равномерной. При последующем росте ошибки неравномерность гистограммы также всё более увеличивается. При этом в разнице изображений появляется текстура, аналогичная текстуре, присутствующей на несжатом изображении, что говорит о том, что часть пространственной информации была потеряна при сжатии. Сами сжатые изображения при этом остаются визуально практически неотличимыми от оригинального, а отличия заметны только при переходе на пиксельный уровень. Также при увеличении допустимой ошибки начинает проявляться эффект занижения значений в сжатом изображении, который усиливается с ростом заданной ошибки. Такой эффект наблюдается при сжатии алгоритмом LERC изображений, в которых присутствуют сравнительно однородные области, значения в которых варьируются меньше, чем на заданную ошибку.

Таким образом, несмотря на то, что с точки зрения сохранения радиометрических свойств изображения допустимая ошибка в 2,5–3 % не является существенной, она уже искажает пространственную текстуру оригинального изображения, поэтому именно критерий сохранения пространственной структуры изображений является определяющим при выборе максимальной ошибки при использовании алгоритма сжатия LERC.

Методика выбора допустимых ошибок при использовании алгоритма LERC для разных типов данных

Для практического использования рассматриваемого алгоритма сжатия необходимо выработать критерии и методику выбора допустимой максимальной ошибки LERC, при которой вносимые текстурные искажения не будут существенно влиять на результаты обработки и анализа сжатых изображений. Для создания такой методики были рассмотрены подходы, основанные на методах сегментации и классификации изображений, а также подходы, построенные на расчёте статистических характеристик полученных изображений.

Результаты работы методов сегментации и классификации (Oti et al., 2021), особенно без обучения, неустойчивы и сильно зависят от обрабатываемых данных, что делает эти методы непригодными для получения универсальной метрики оценки искажения пространственной структуры.

Поэтому были выбраны статистические критерии для оценки искажений, вносимых в пространственную структуру изображений. Как было показано выше, при появлении пространственной структуры в разности между исходным и сжатым изображениями их гистограммы «перекашиваются» вправо, поэтому в качестве первого критерия было выбрано среднее значение по разнице. Однако такой показатель характеризует только изображение в целом. В качестве более детального способа выявления структуры в разнице между изображениями был использован индекс автокорреляции изображения со сдвигом 1. Так как

размеры изображений в пикселях, как правило, достаточно велики, искажения, связанные с использованием одномерного массива значений вместо двумерного, можно считать несущественными. 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.

Для проведения экспериментов использовалась описанная выше выборка восстановленных композитов индекса NDVI (диапазон значений от 0 до 10 000), полученных по данным прибора MSI, установленного на спутниках серии Sentinel-2. Для каждого сжатого изображения были рассчитаны средние значения по разности, индексы автокорреляции, а также степень сжатия. Потом для каждого из значений максимальных ошибок сжатия получены средние (агрегированные) значения этих показателей. Для рассчитанных агрегированных значений была построена корреляционная матрица, которая показала очень высокую корреляцию (около 0,98) между средним по разности и автокорреляцией со сдвигом 1. Это позволяет предположить, что среднее значение по разности между исходными и сжатыми изображениями так же хорошо характеризует вносимые пространственные искажения, как и индекс автокорреляции со сдвигом 1. Примерно такая же высокая степень корреляции была обнаружена у рассматриваемых показателей и со степенью сжатия изображений, вычисляемой как отношение размера исходного изображения к размеру сжатого. Поскольку все выбранные показатели хорошо коррелируют, для оценки пространственных искажений в дальнейшем достаточно выбрать какой-либо один из них, например среднее по разности, как наиболее простое в вычислении. На *рис. 4* приведена двумерная гистограмма зависимости между индексом автокорреляции со сдвигом 1 (представлен по оси *X*) и величиной среднего значения по разности между исходным и сжатыми изображениями (отложена по оси *Y*), построенная для наборов данных с максимальной ошибкой сжатия в 0,5, 1, 1,5, 2 и 3 %. На представленной гистограмме видно, что для малых значений автокорреляции и среднего, которые как раз представляют основной интерес, эти два показателя действительно хорошо коррелированы между собой, на больших же значениях параметров корреляция падает.

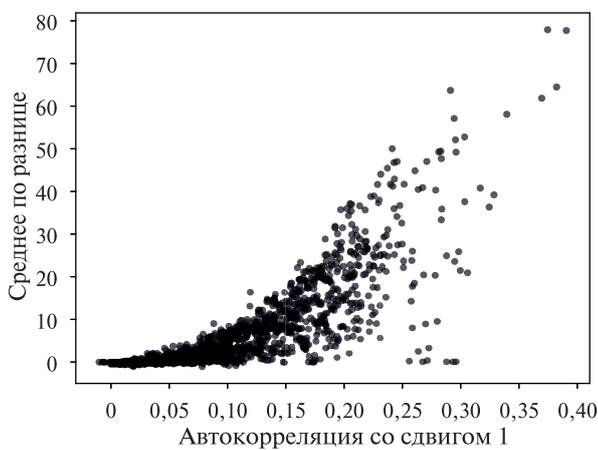


Рис. 4. Гистограмма корреляции индекса автокорреляции (ось *X*) и среднего по разности изображений (ось *Y*)

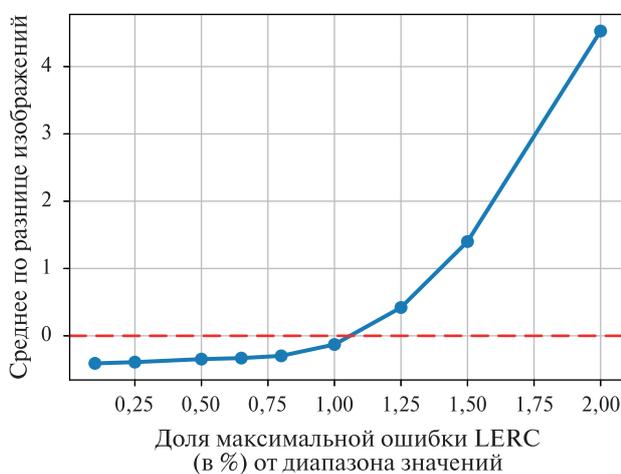


Рис. 5. Зависимость средних значений по разности изображений от максимальной ошибки LERC на основе данных NDVI, полученных по измерениям MSI

На *рис. 5* представлена зависимость агрегированного по тестовой выборке среднего значения по разности изображений от максимальной ошибки LERC. Видно, что в нашем случае 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.

На графиках хорошо видно, что после пересечения средним по разности нулевого значения скорость роста среднего в зависимости от ошибки существенно увеличивается, поэтому в качестве порогового значения для допустимой ошибки LERC было выбрано значение, при котором агрегированное среднее по разности ближе всего к нулю. Для рассматриваемого тестового набора это значение составляет примерно 1 % от исходного диапазона значений в изображениях.

Для подтверждения выбранного порогового значения нами было исследовано поведение гистограмм и агрегированных средних по разности для значений максимальной ошибки сжатия в диапазоне от 0,25 до 1,5 %, представленных на *рис. 6*. Ошибка дана в процентах от исходного диапазона, средние — в абсолютных значениях. Нулевые значения были выколоты для большей детальности приведённых гистограмм.

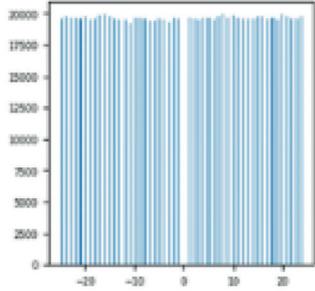
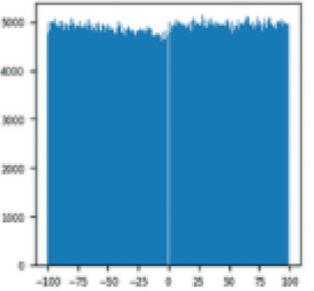
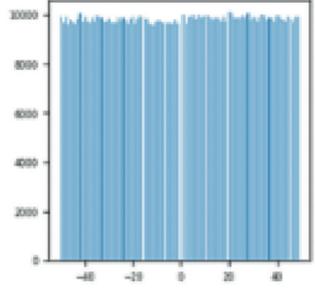
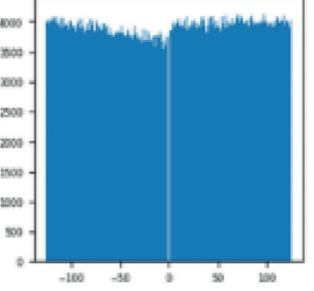
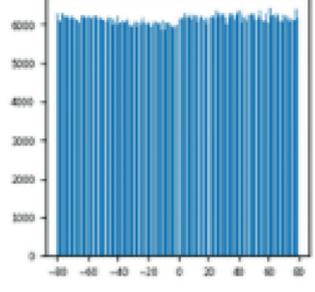
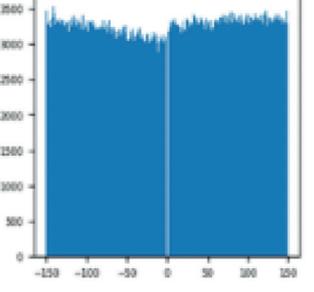
Ошибка LERC, %	Гистограмма разности	Среднее	Ошибка LERC, %	Гистограмма разности	Среднее
0,25		-0,40	1,00		-0,1
0,50		-0,35	1,25		0,4
0,80		-0,30	1,50		1,4

Рис. 6. Поведение гистограмм разности и среднего значения для ошибок LERC в диапазоне 0,25–1,5 %

Хорошо видно, что при малых ошибках сжатия гистограммы значений в разности остаются стабильными и отражают распределение значений, близкое к шуму, среднее значение по разности также остаётся стабильным. После превышения выбранного порога в 1 % гистограмма перестаёт быть стабильной (см. также *рис. 3*), средние значения по разности также

резко начинают расти (для ошибки в 1,5 % среднее по разности составляет 1,4, для ошибки в 2 % — 4,5). 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).

Для дополнительного обоснования такого выбора была также проанализирована гистограмма распределения средних ошибок по разнице значений для всех файлов в выборке для разных ошибок LERC, приведённая на рис. 7. При ошибках больше чем 1 % резко увеличивается дисперсия распределения. Это означает, что для значительной части изображений в тестовой выборке пространственные искажения существенно выше, чем в среднем, и выбранный нами порог корректен.

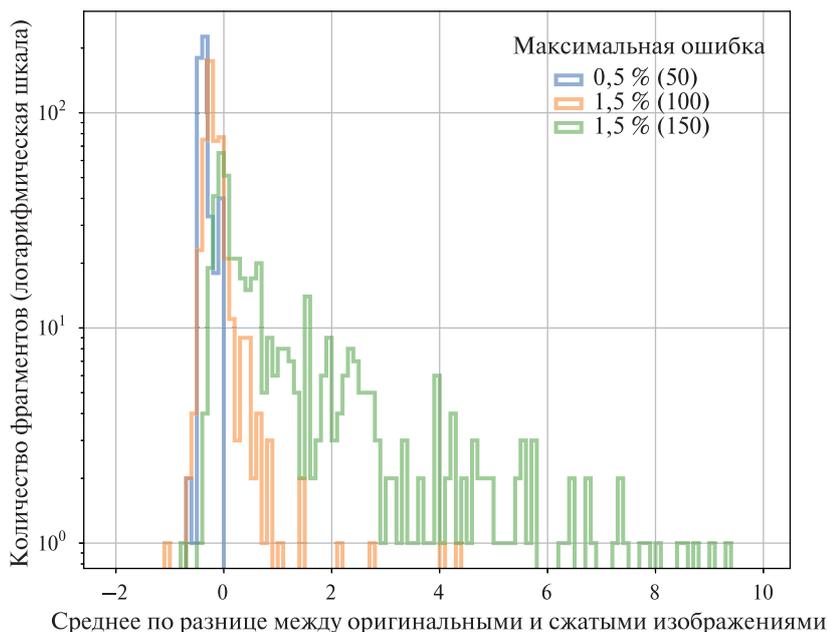


Рис. 7. Распределение средних ошибок по разнице изображений в выборке для разных допустимых ошибок LERC

Для выведения методики также необходимо было оценить корректность нашего подхода и достаточность созданной выборки из 500 элементов для получения распределения индекса автокорреляции. Для этого была использована упрощённая формула для оценки минимального размера выборки для достижения требуемой точности при определении среднего по ней значения:

$$MSS = (2S/P)^2,$$

где S — стандартное отклонение; P — требуемая точность.

Из этой формулы можно получить выражение для расчёта точности для известного размера выборки:

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

Для распределения средних значений, соответствующих выявленной пороговой ошибке в 1 % (абсолютное значение ошибки 100), дисперсия среднего по разнице изображений примерно равна 0,4. Согласно приведённой формуле выборка из 500 элементов позволяет получить среднее по этой величине с точностью лучше, чем 0,036. Для достижения же допустимой точности в 0,05 достаточно и вдвое меньшей выборки. Можно предположить, что и для других типов спутниковых изображений оценки необходимого размера выборки будут похожими.

В итоге по результатам проведённого исследования была выработана следующая методика для определения порогового значения ошибки LERC, позволяющего получить максимально возможное сжатие данных без потери радиометрических и структурных свойств изображения. Методика включает в себя следующие этапы:

1. Подготовка случайной по времени и географическому расположению выборки спутниковых изображений анализируемого типа, состоящей из нескольких сотен экземпляров данных.
2. Получение набора сжатых алгоритмом LERC изображений для ряда допустимых ошибок LERC в диапазоне от 0,5 до меньшего из половины диапазона радиометрической точности данных или 2 % от диапазона значений в изображениях для нескольких экземпляров из выборки.
3. Получение зависимости агрегированного среднего по разности между оригинальными и сжатыми изображениями от максимальной ошибки LERC и выбор значения, при котором эта величина максимально близка к нулю, в качестве порогового.
4. Повтор шагов 2 и 3 для всей выборки для диапазона максимальных ошибок сжатия $\pm 0,5\%$ от полученного порогового значения для уточнения найденного порога.

Предложенная методика была применена для получения порогового значения сжатия для восстановленного композита канала RED по данным спутников Sentinel-2. Диапазон используемых значений для этого продукта существенно уже (около 2500), а установленное пороговое значение ошибки LERC равно примерно 1,2 % (около 30). При этом достигается относительная степень сжатия около 3,32. На *рис. 8* приведена зависимость агрегированного среднего значения по разнице между оригинальными и сжатыми изображениями для восстановленного композита канала RED по данным спутников Sentinel-2.

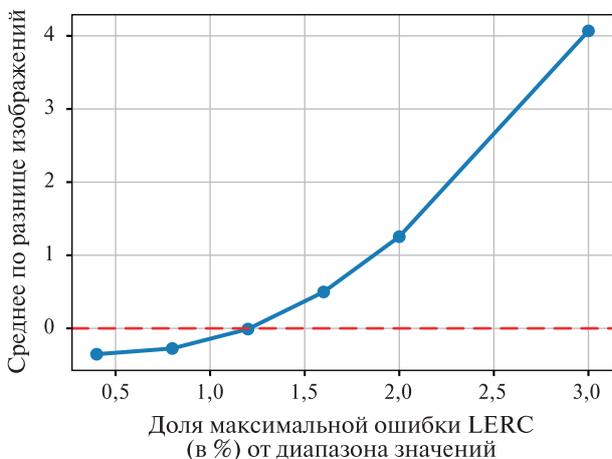


Рис. 8. Зависимость агрегированных средних значений по разнице изображений от максимальной ошибки LERC для данных Sentinel-2 RED

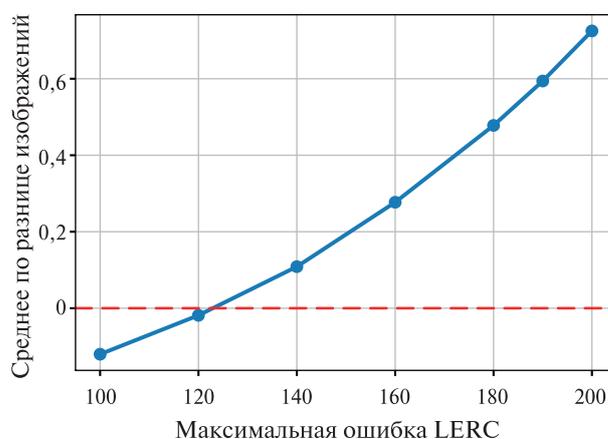


Рис. 9. Зависимость агрегированных средних значений по разнице изображений от максимальной ошибки LERC для данных MODIS NDVI

Также методика была применена для восстановленных композитов индекса NDVI по данным прибора MODIS, реализованных в том же диапазоне значений, что и аналогичный продукт по данным спутников Sentinel-2. На *рис. 9* приведена аналогичная зависимость для выборки (график зависимости средних по разности от максимальной ошибки сжатия). Полученное пороговое значение ошибки LERC равно примерно 1,2 % от диапазона значений, при этом достигается относительная степень сжатия данных около 2,9.

Результаты исследования показывают, что использование алгоритма сжатия с потерей качества LERC, гарантирующего максимальную величину ошибки при сжатии, может быть допустимым при работе с данными ДЗЗ не только при задачах визуализации данных, но и в

задачах, связанных с их обработкой и анализом. Корректно подобранные пороговые значения максимальной ошибки позволяют проводить сжатие без потери радиометрических и структурных свойств изображения, вместе с тем обеспечивая существенный, до 3–4 раз, выигрыш в степени сжатия по сравнению с традиционными алгоритмами сжатия данных без потерь, что может привести как к существенной экономии аппаратных ресурсов хранения данных, так и к снижению нагрузки на сети передачи данных и времени обработки данных за счёт уменьшения накладных расходов.

Заключение

Представлен анализ применимости алгоритма сжатия изображений с потерями LERC для архивации данных ДЗЗ с точки зрения сохранения радиометрических и структурных свойств изображения. Было установлено, что ключевым критерием, больше всего влияющим на выбор порогового значения ошибки для этого алгоритма, является сохранение пространственной структуры спутниковых изображений. Для определения допустимого порогового значения ошибки была разработана методика, основанная на статистических критериях для оценки вносимых в сжатые изображения искажений. Результаты применения этой методики для созданных тестовых выборок показали, что использование алгоритма LERC с полученными значениями максимальной ошибки позволяет сократить объём данных в архиве более чем в три раза относительно данных, сжатых при помощи алгоритма DEFLATE, не потеряв при этом радиометрических и структурных свойств данных, а сам алгоритм LERC имеет широкие перспективы применения для реальных задач обработки и анализа данных при условии соблюдения критериев, обеспечивающих сохранность свойств изображения. Так, его применение к накопленным в ЦКП «ИКИ-Мониторинг» архивам исходных данных спутников Sentinel-2 уровней обработки L1C и L2A, объём которых превышает 3 Пб, потенциально может уменьшить эту величину примерно на 2 Пб.

Исследование было выполнено на базе использования возможностей Центра коллективного пользования системами архивации, обработки и анализа данных спутниковых наблюдений «ИКИ-Мониторинг» и при поддержке темы Министерства науки и высшего образования Российской Федерации «Большие данные в космических исследованиях: астрофизика, солнечная система, геосфера» (№ 122042500019-6).

Литература

1. Луян Е. А., Прошин А. А., Бурцев М. А., Балашов И. В., Барталев С. А., Ефремов В. Ю., Кашицкий А. В., Мазуров А. А., Матвеев А. М., Суднева О. А., Сычугов И. Г., Толпин В. А., Уваров И. А. Центр коллективного пользования системами архивации, обработки и анализа спутниковых данных ИКИ РАН для решения задач изучения и мониторинга окружающей среды // Современные проблемы дистанционного зондирования Земли из космоса. 2015. Т. 12. № 5. С. 263–284.
2. Луян Е. А., Прошин А. А., Бурцев М. А. и др. Опыт эксплуатации и развития центра коллективного пользования системами архивации, обработки и анализа спутниковых данных (ЦКП «ИКИ-Мониторинг») // Современные проблемы дистанционного зондирования Земли из космоса. 2019. Т. 16. № 3. С. 151–170. DOI: 10.21046/2070-7401-2019-16-3-151-170.
3. Миклашевич Т. С., Барталев С. А., Плотников Д. Е. Интерполяционный алгоритм восстановления длинных временных рядов данных спутниковых наблюдений растительного покрова // Современные проблемы дистанционного зондирования Земли из космоса. 2019. Т. 16. № 6. С. 143–154. DOI: 10.21046/2070-7401-2019-16-6-143-154.
4. Becker P., Plesea L., Maurer T. Cloud optimized image format and compression // Intern. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2015. V. 40. P. 613–615. DOI: 10.5194/isprsarchives-XL-7-W3-613-2015.
5. Becker P., Plesea L., Maurer T. Optimizing cloud based image storage, dissemination and processing through use of MRF and LERC // Intern. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2016. V. 41. P. 201–203. DOI: 10.5194/isprsarchives-XLI-B4-201-2016.

6. Oswal S., Singh A., Kumari K. Deflate compression algorithm // Intern. J. Engineering Research and General Science. 2016. V. 4. No. 1. P. 430–436.
7. Oti E. U., Olusola M. O., Eze F. C., Enogwe S. U. Comprehensive review of K-Means clustering algorithms // Intern. J. Advances in Scientific Research and Engineering. 2021. V. 7. Iss. 8. P. 64–68. DOI: 10.31695/IJASRE.2021.34050.

Features of LERC image compression algorithm application in Earth remote sensing data archiving

A. A. Proshin, E. A. Loupian, M. A. Burtsev

Space Research Institute RAS, Moscow 117997, Russia

E-mail: andry@smis.iki.rssi.ru

Recent decades have provided an almost exponential growth of remote sensing data volumes. This is due both to the increase in the number of operational, including open-access, satellite systems and to the improvement of performance of imaging systems and the growth in the number of available information products. Therefore, the search for new approaches for more efficient remote sensing data storage, including through the improvement of their compression mechanisms that allow preserving the required data quality and providing high speed access to them, including the use of lossy compression algorithms, remains one of the urgent tasks. This paper is devoted to analyzing the possibility of using the LERC (Limited Error Raster Compression) lossy compression algorithm, which allows setting the maximum error in a pixel, for compression of various types of remote sensing satellite data. The paper considers the influence of a set maximum error on both the degree of data compression and the features of introduced distortions both in the radiometric properties of the data, the level of which the algorithm allows to control, and in the spatial characteristics of the compressed images. Special focus is on the analysis of image texture distortions depending on the level of allowable error used in compression. In order to determine the maximum allowable compression error that does not lead to the loss of either radiometric or structural properties of the image, a technique based on quantitative assessment of changes observed in image texture at different levels of maximum error is proposed, and the results of application of the technique to different types of remote sensing data acquired in the visible range and products obtained from their processing are presented.

Keywords: satellite data, Earth remote sensing, satellite data archives, efficient data compression, satellite data access systems

Accepted: 21.10.2024

DOI: 10.21046/2070-7401-2024-21-6-130-142

References

1. Loupian E. A., Proshin A. A., Burtsev M. A., Balashov I. V., Bartalev S. A., Efremov V. Yu., Kashnitskii A. V., Mazurov A. A., Matveev A. M., Sudneva O. A., Sychugov I. G., Tolpin V. A., Uvarov I. A., IKI-Monitoring center for collective use of systems for archiving, processing and analyzing satellite data for environment research and analysis problems, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2015, Vol. 12, No. 5, pp. 263–284 (in Russian).
2. Loupian E. A., Proshin A. A., Burtsev M. A. et al., Experience of development and operation of the IKI-Monitoring center for collective use of systems for archiving, processing and analyzing satellite data, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2019, Vol. 16, No. 3, pp. 151–170 (in Russian), DOI: 10.21046/2070-7401-2019-16-3-151-170.
3. Miklashevich T. S., Bartalev S. A., Plotnikov D. E., Interpolation algorithm for reconstruction of vegetation cover satellite observations long time series, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2019, Vol. 16, No. 6, pp. 143–154 (in Russian), DOI: 10.21046/2070-7401-2019-16-6-143-154.

4. Becker P., Plesea L., Maurer T., Cloud optimized image format and compression, *Intern. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2015, Vol. 40, pp. 613–615, DOI: 10.5194/isprsarchives-XL-7-W3-613-2015.
5. Becker P., Plesea L., Maurer T., Optimizing cloud based image storage, dissemination and processing through use of MRF and LERC, *Intern. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016, Vol. 41, pp. 201–203, DOI: 10.5194/isprsarchives-XLI-B4-201-2016.
6. Oswal S., Singh A., Kumari K., Deflate compression algorithm, *Intern. J. Engineering Research and General Science*, 2016, Vol. 4, No. 1, pp. 430–436.
7. Oti E. U., Olusola M. O., Eze F. C., Enogwe S. U., Comprehensive review of K-Means clustering algorithms, *Intern. J. Advances in Scientific Research and Engineering*, 2021, Vol. 7, Iss. 8, pp. 64–68, DOI: 10.31695/IJASRE.2021.34050.