Альтернативный подход к сжатию и хранению исходных временных рядов.

Махмудов Е.Р., Яблочкина К.А.

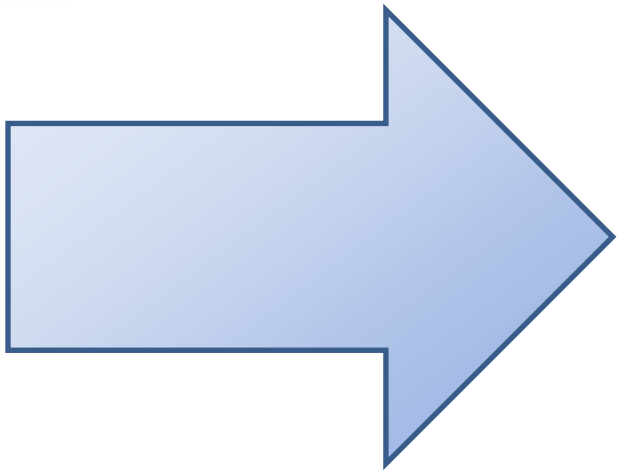
Камчатский филиал Геофизической службы РАН, г. Петропавловск-Камчатский

mer@emsd.ru

Пример сжатия текста

364
96
-89
-387
-419
-255
-298
-516
-615
-435
16
376
632
527
184
-186
-506
-785
-723
-401
43
207
316
387
342
524
677
628
533
504
465
494
516
465
165
-59

Исходный текст = 152 байта



ВяьbR┼_ў_мЇ
Б°Pvт6l/M;▒╢иT▄EXk┘ оШ,(p→═@┼ └7a╛ "ъHfEMPь
q$◄g ┗▀╤аця$z♠↔Л"Q'C] ╚ЫI°П╛ wH ╒}|Ц↑■

Сжатый текст = 95 байт

Разница 60%

Алгоритмы сжатия

•deflate - использует комбинацию алгоритма LZ77 и алгоритма Хаффмана, применяется в программах WinZip, gzip. Утвержден в 1996г.

•Bzip2 - реализация алгоритма Барроуза — Уилера блочно-сортировочного сжатия, применяется в программе bzip2 и тд. Утвержден в конце 2000г.

•LZMA - улучшенный алгоритм сжатия LZ77, дополненный алгоритмом интервального кодирования, применяется в программе 7-Zip. Разработан в первой половине 2000ых.

PPMd – алгоритм основан на контекстном моделировании и предсказании, используется в программах 7-Zip, WinZip и т.д. Разработан в 1980ых г.

Сравнение алгоритмов сжатия

	Deflate	Bzip2	LZMA	PPMd
Коэффициент сжатия	3,19	4,44	3,89	4,78
Коэффициент сжатия, %	100	139	121	150
Время сжатия, с	103	133,1	107,6	29,7
Время сжатия, %	100	77,3	95,7	346
Время распаковки, с	0,76	4,4	1,6	30,5
Время распаковки, %	100	17,2	47,5	2,5
Оценка распространенности для разных языков программирования	5	4	3	1

Сравнение алгоритмов сжатия. Выводы.

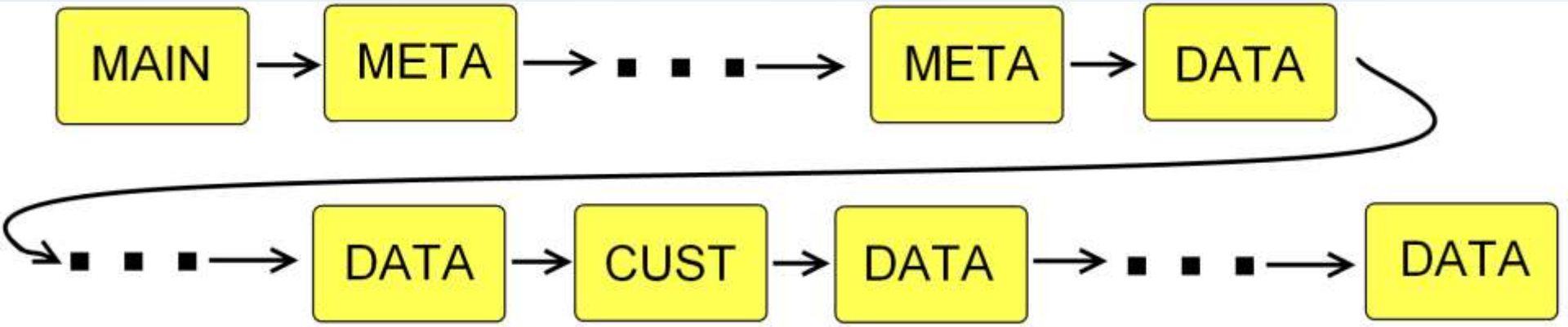
- Для всех алгоритмов параметры сильно зависят от входящих данных и не могут быть оценены заранее
- алгоритм PPMd не подходит для подобных задач ввиду большого времени распаковки, поэтому исключается из рассмотрения
- алгоритм deflate однозначно лучше по скорости распаковки
- алгоритм Bzip2 имеет лучший коэффициент сжатия
- алгоритм LZMA занимает промежуточное положение, имея хорошие показатели.
- Несмотря на то, что алгоритм LZMA по характеристикам выглядит оптимальным, на сегодняшний день предпочтительнее использование алгоритма bzip2, однако общая тенденция к расширению применения алгоритма LZMA позволит в будущем говорить об его рекомендации к использованию.

TCTiSe

(от Text Compressed Time Series в переводе "сжатые текстовые временные серии", произносится как ТиСиТайз)

Формат данных общего назначения для хранения значений временных рядов. Имеет фиксированную блочную структуру для служебных блоков, блоки данных переменной длины и динамическую структуру у свободно определенных блоков.

Схема структуры формата



Первым должен следовать основной блок MAIN, затем должны располагаться информационные блоки META. Следующие блоки с данными DATA и блоки расширения CUST могут чередоваться.

Блок MAIN

Содержит системную информацию о формате данных, общую информацию о регистрации.

Наименование	Тип данных	Значение	Описание
Идентификатор типа блока	строка 10с	"BLOCK MAIN"	Строка с фиксированным значением
Название формата файла	строка 10с	" TCTISE"	Строка с фиксированным значением
Вариация и версия формата	строка 4с		Два первых символа являются элементами латинского алфавита и означают вариацию формата, два последних являются цифрами и означают версию формата файла. Для каждой пары символов действует правило дополнения пробелами слева. Например " A 0"
Порядок байтов	строка 1с	"<", ">"	Порядок следования байтов бинарных данных, little-endian соответствует "<", big-endian соответствует ">". Рекомендуется использовать big-endian порядок.
Датавремя начала	строка 19с		Используется читабельный вид представления даты и времени, например "2012-06-13 12:00:00"
Координаты	строка 22с		Гео координаты места регистрации, представленные двумя дробными числами, разделенными пробелом. Широта представлена 10 символами, долгота 11. Для каждой координаты действует правило левого дополнения. Пример "-12.123456 -179.123456", " 53.066756 158.607441"
Комментарий	строка 250с		Текстовый комментарий свободного содержания
Зарезервировано	10 байт		Зарезервированные байты

Блок META

Содержит информацию о канале регистрирующей станции, о характеристиках измеряемого параметра.

Наименование	Тип данных	Значение	Описание
Идентификатор типа блока	строка 10с	"BLOCK META"	Строка с фиксированным значением
Название станции	строка 5с		Например " KLY"
Название канала	строка 5с		Например "WINDSP"
Название сети	строка 5с		Например " N1"
Частота дискретизации	uint 4b		Количество измерений в секунду, например 100
Период дискретизации	uint 4b		Время между соседними измерениями в микросекундах, например если значение 10мс (соответствует 100Гц), то значение будет 10000
Метод сжатия	строка 10с		Например " bzip2"
Тип упаковываемого значения	строка 10с		Как интерпретировать текстовые распакованные значения в типах данных C
Измеряемый параметр	строка 5с		Например " m/s2"
Название датчика	строка 20с		Например " Hakusan Si102"

Блок DATA

Содержит контрольные метки и упакованные значения временного ряда.

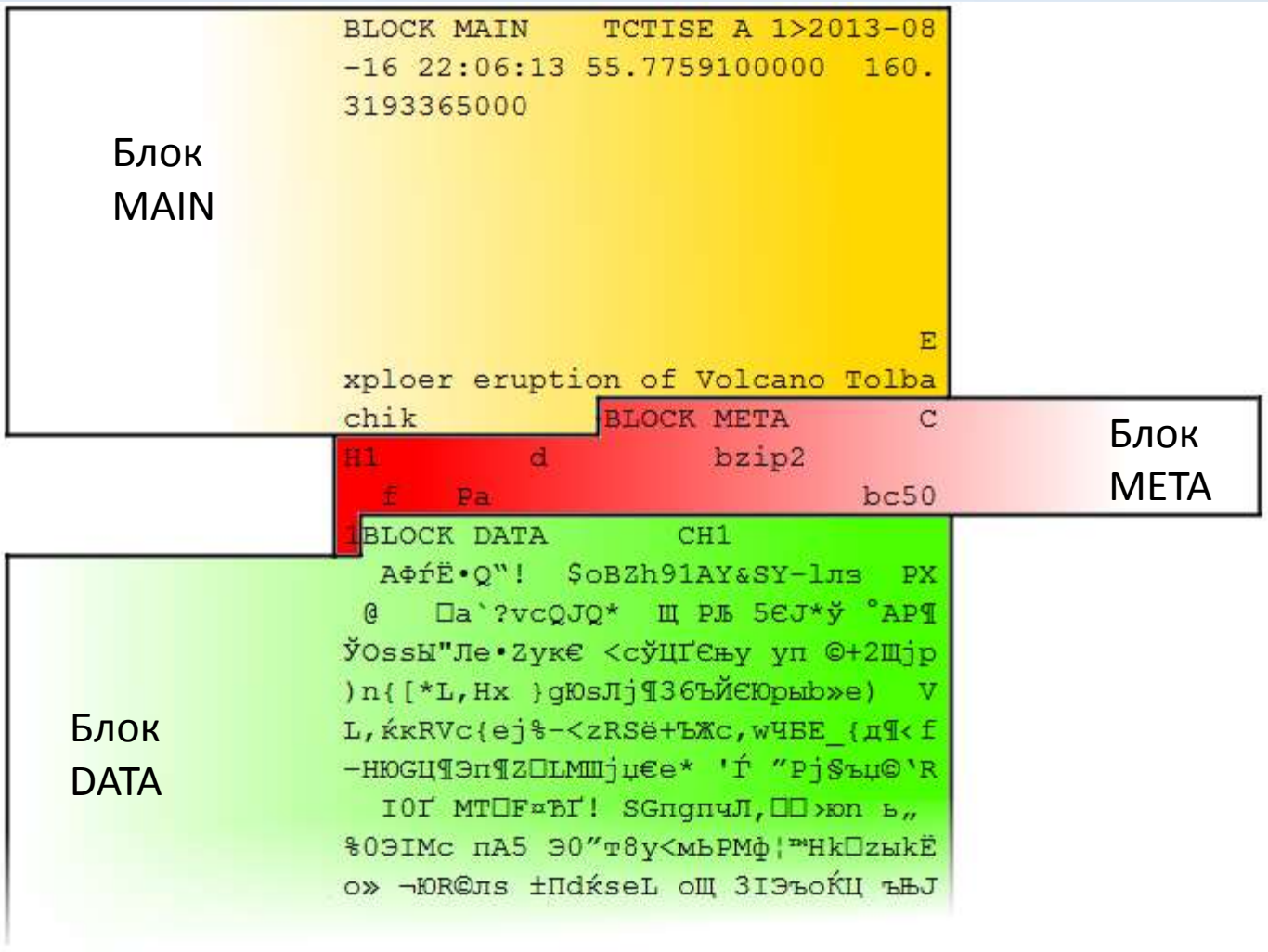
Наименование	Тип данных	Значение	Описание
Идентификатор типа блока	строка 10с	"BLOCK DATA"	Строка с фиксированным значением
Название станции	строка 5с		Например " KLY"
Название канала	строка 5с		Например "WINDSP"
Название сети	строка 5с		Например " N1"
ID блока глобальный	uint 4b		Последовательный номер блока от момента начала аппаратной регистрации
ID блока канальный	uint 4b		Последовательный номер блока для канала от момента начала аппаратной регистрации
Датавремя	double 8b		Отметка даты и времени начала блока, выраженное дробным числом секунд с начала 1970 года,
			является возвращаемым значением функции C time()
Размер данных	uint 4b		Длина запакованной строки с данными
Данные	строка		Строка переменной длины содержащая запакованный временной ряд

Блок CUST

Блок-расширение, структура определяется свободно, при соблюдении базовых полей. Может быть использован для хранения второстепенных регистрируемых данных, являться уведомлением о происходящих сбоях и т.п.

Наименование	Тип данных	Значение	Описание
Идентификатор типа блока	строка 10с	"BLOCK CUST"	Строка с фиксированным значением
Идентификатор расширения	строка 32с		Уникальный идентификатор определяющий тип CUSTOM блока,должен использоваться в качестве ключа при регистрации в пользовательском приложении для извлечения стукруры данных из блока. Рекомендуется использовать md5 хеш сумму строки подробно описывающей предназначение блока.
Размер блока	uint 4b		Размер последующего блока данных в байтах

Пример содержания файла TCTiSe

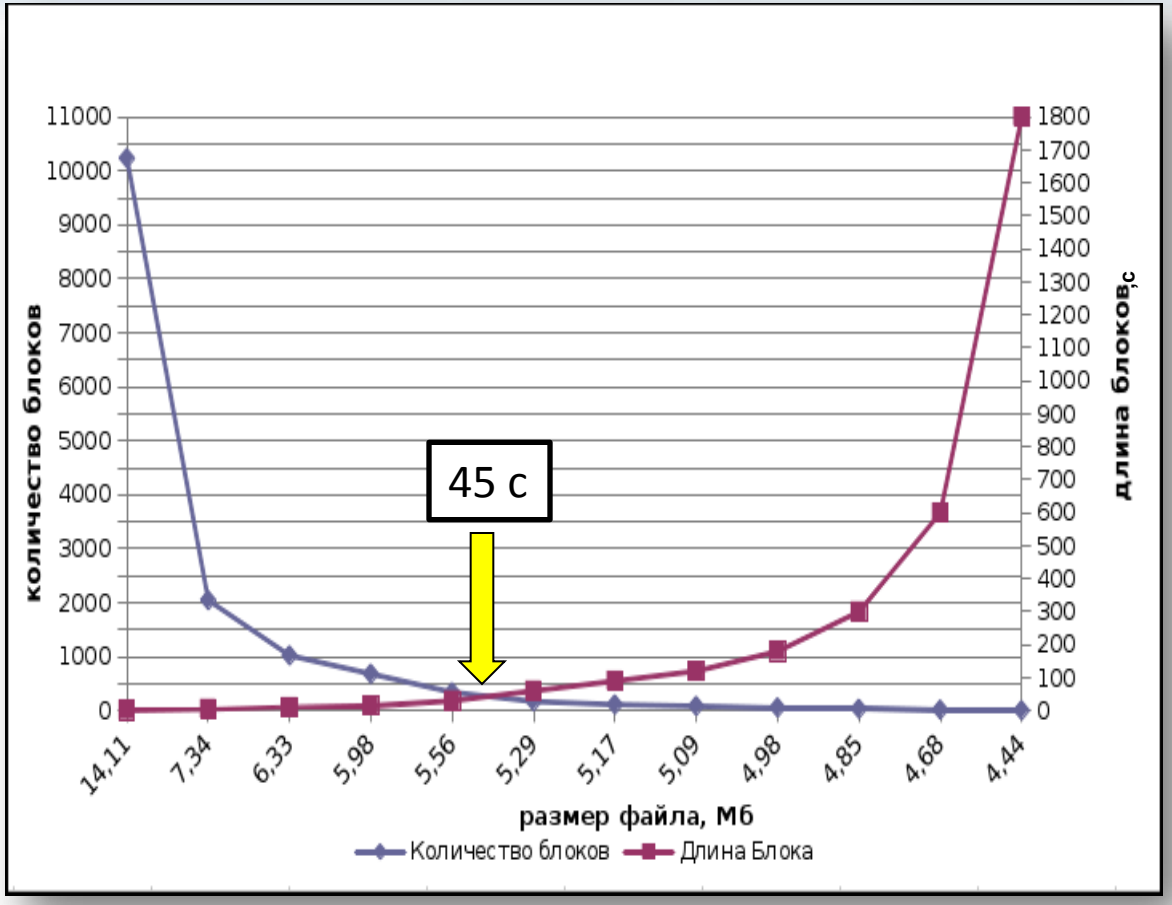


Сравнение форматов данных

Формат / Сигнал	ASCII	SEED	WIN	TCTiSe_Gzip	TCTiSe_Bzip2	TCTiSe_LZMA	$\frac{Format}{TCTiSe}\cdot 100\%$
Сейсмика тихая	42,2	6,56	-	6,3	4,5	5,73	140
Сейсмика шумная	56,7	15,0	-	19,4	15,4	17,6	97
Инфразвук	50,2	13	-	17,9	14,5	15,0	89
Наклономер	5,17	-	0,815	0,629	0,499	0,565	163

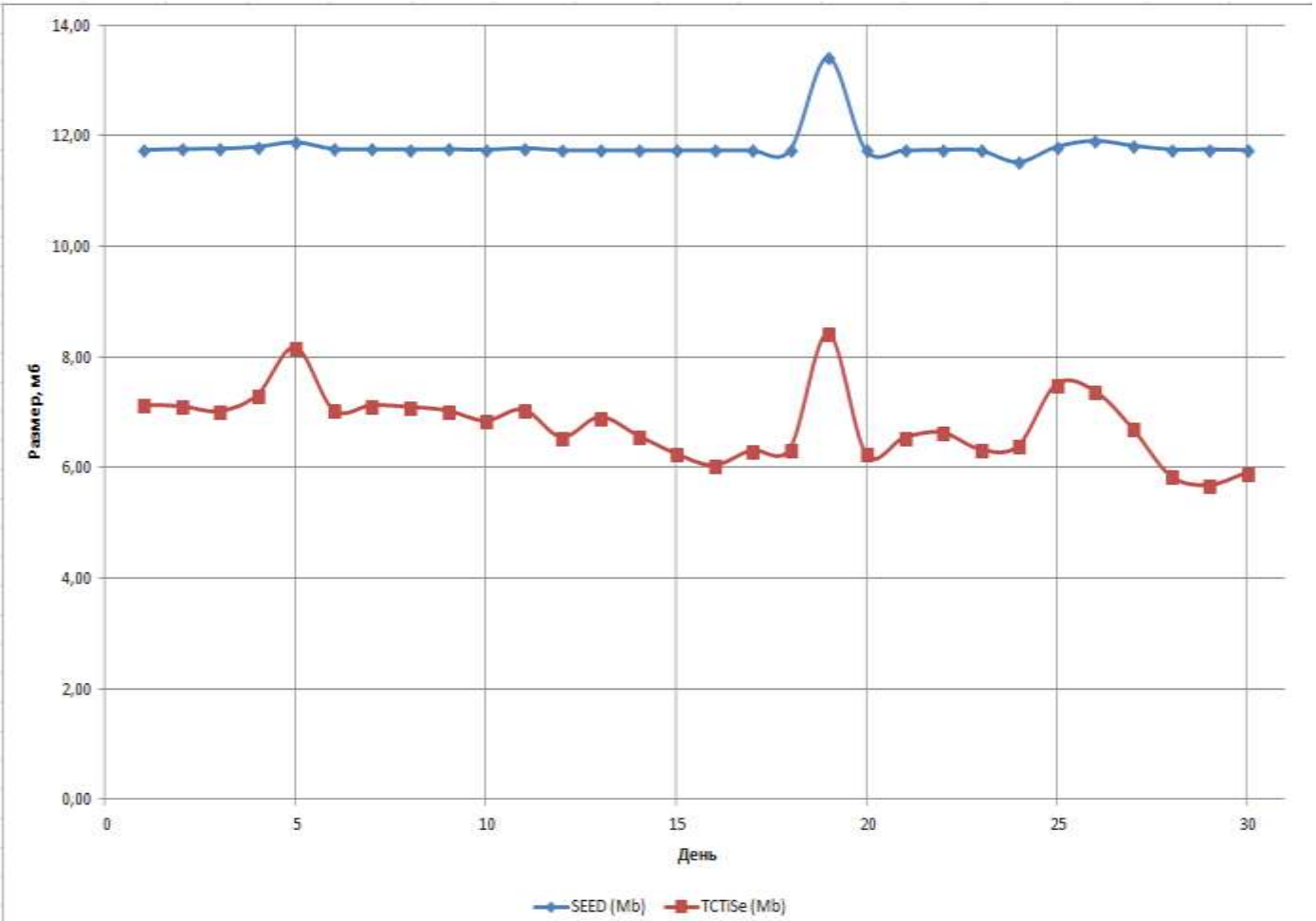
Размер файла форматов данных при различных входных данных. Размеры указаны в Мб. Последняя колонка указывает процентное соотношение формата SEED/WIN и лучшего показателя TCTiSe.

Зависимость размера данных от длины буфера



Типовой график зависимости количества и длины блоков от размера файла, пересечение линий которых, дает представление об оптимальной длине блока в 45 секунд, **однако для удобства рекомендовано использование 30 или 60 секундного блока**.

Сравнительный график объема суточных данных по станции BDR (SHZ) за июнь 2013



Итого за месяц:

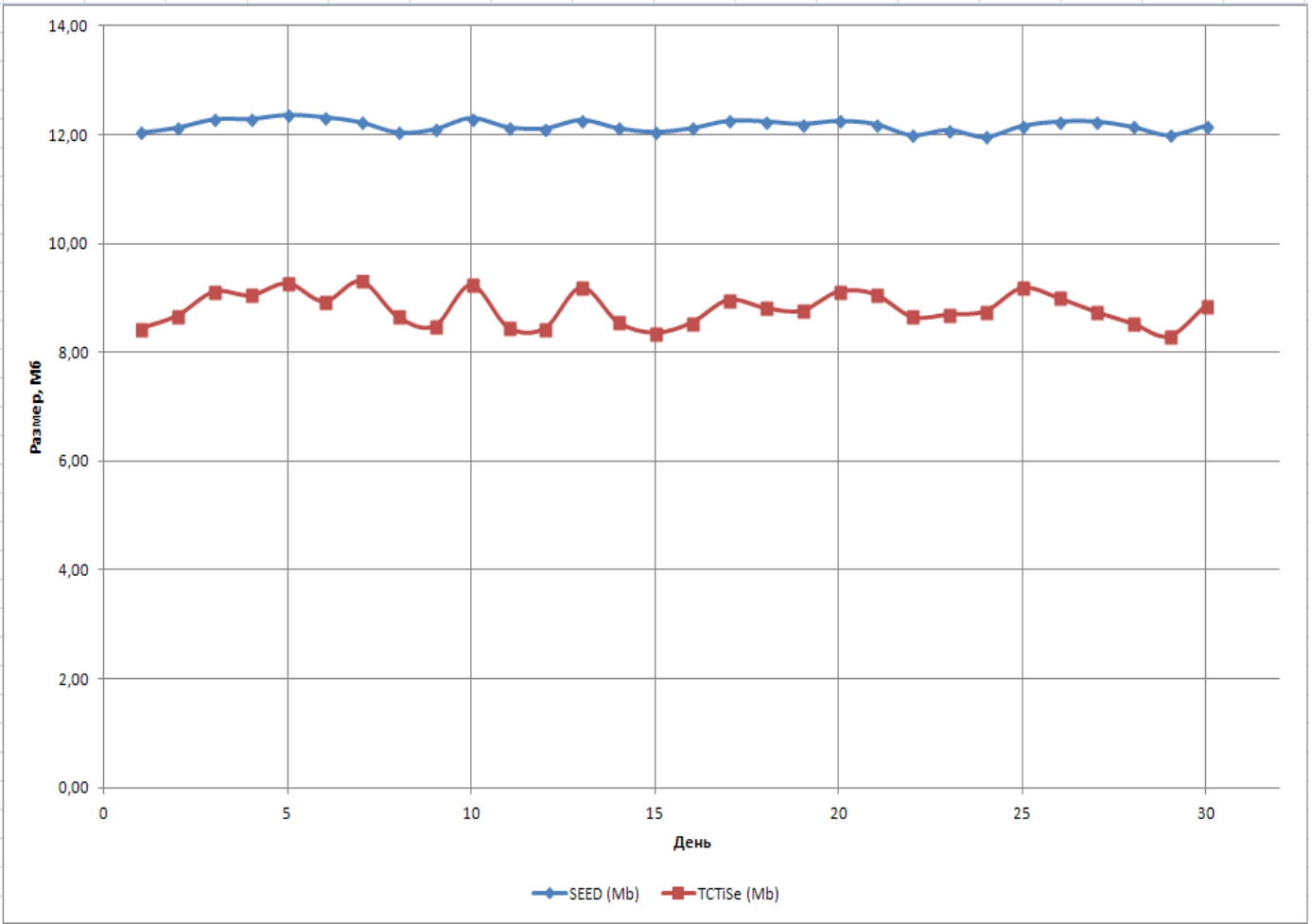
SEED: 354Mb

TCTiSe: 203Mb

Разница 151Mb

или 74%

Сравнительный график объема суточных данных по станции PET (SHZ) за июнь 2013



Итого за месяц:

SEED: 365Mb

TCTiSe: 264Mb

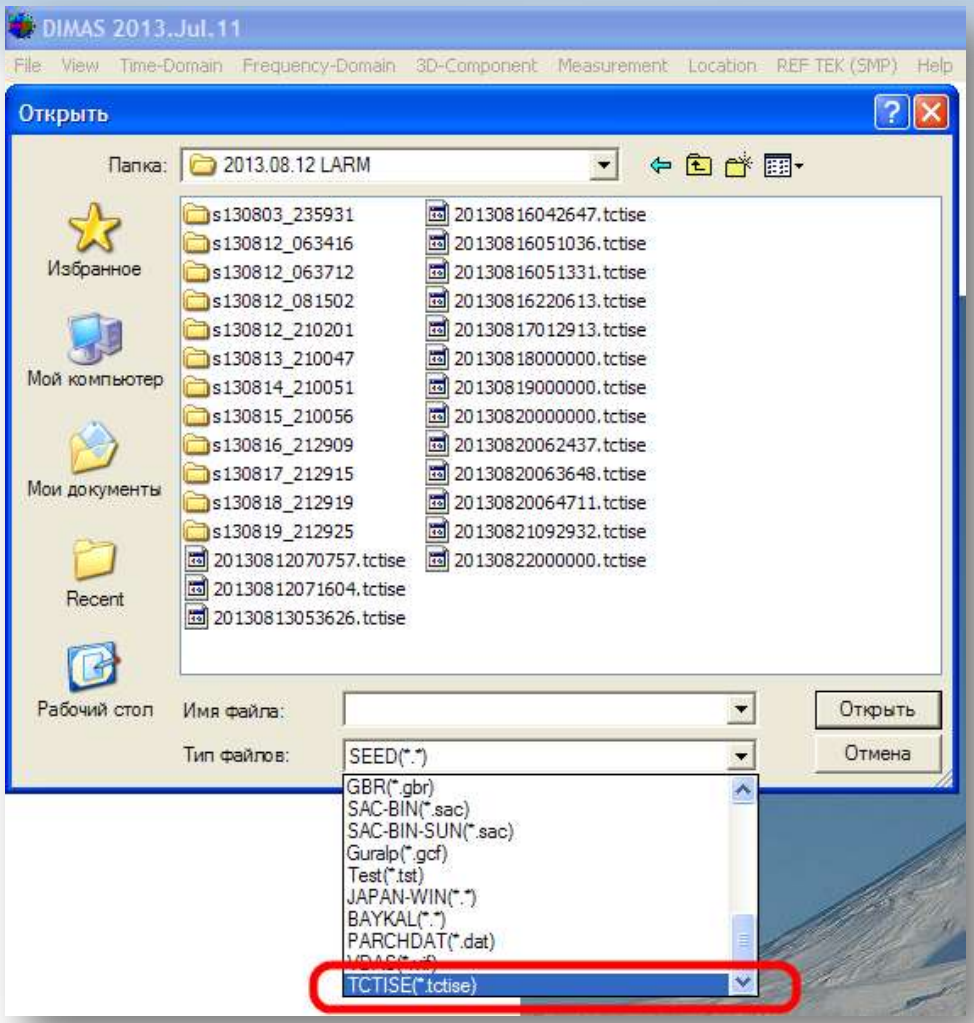
Разница 101Mb

или 38%

Камчатский филиал геофизической службы

Лаборатория акустического и радонового мониторинга

TCTiSe в DIMAS



На текущий момент в DIMAS осуществлена экспериментальная поддержка формата TCTiSe за счет внешней программы-распаковщика (не входит в DIMAS).

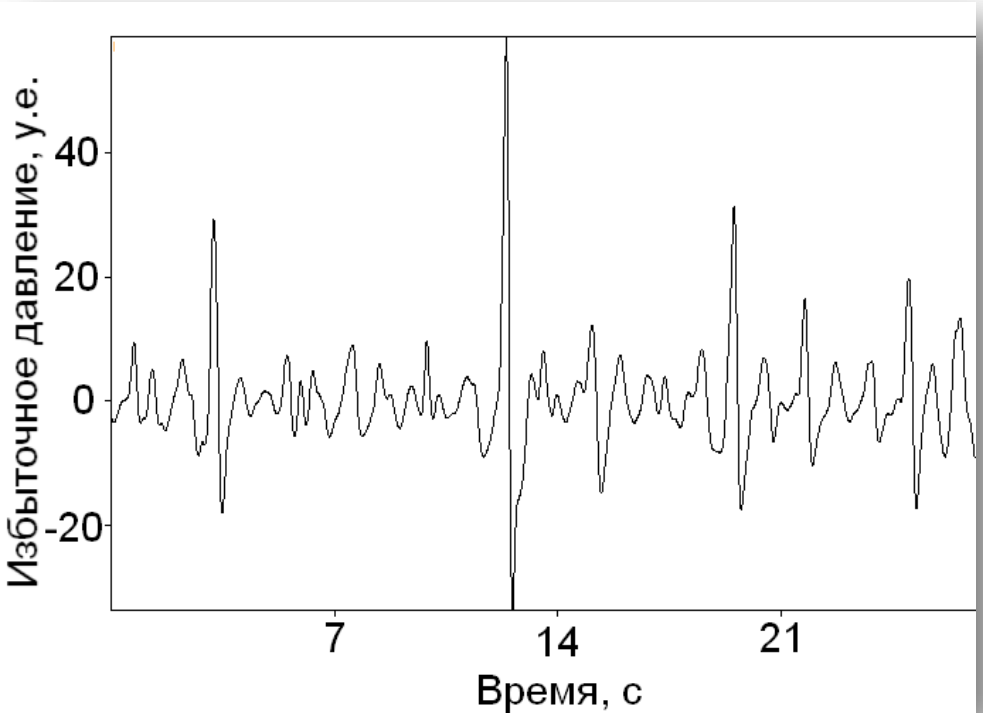
В будущем планируется встроенная поддержка формата.

Камчатский филиал геофизической службы

Лаборатория акустического и радонового мониторинга

Полевые испытания

Формат TCTiSe был успешно опробован при регистрации акустических волн во время извержения прорыва вулкана Плоский Толбачик в августе 2013 года.



Заключение

Отличительные черты формата TCTiSe:

- Крайне простая и ясная структура
- Простая техническая реализация – алгоритмы сжатия широко доступны
- Эффективность сжатия временных рядов (высокая эффективность в ряде случаев)