Real Information Status

Recap of analysis in DRBSD-22 Paper

Total amount of real information M and corresponding compression parameters necessary to save 99%

Variable	M	K for BG_K	P for ZFP_P
FLNS	0.87	2	8
ICEFRAC	6.06	N/A	N/A
LHFLX	3.09	2	8
PRECT	0.61	2	N/A
PSL	2.91	4	12
TAUX	1.33	2	8
TS	3.81	3	12
Q200	2.62	2	8
WSPDSRFAV	1.58	2	8
Z500	3.00	3	12

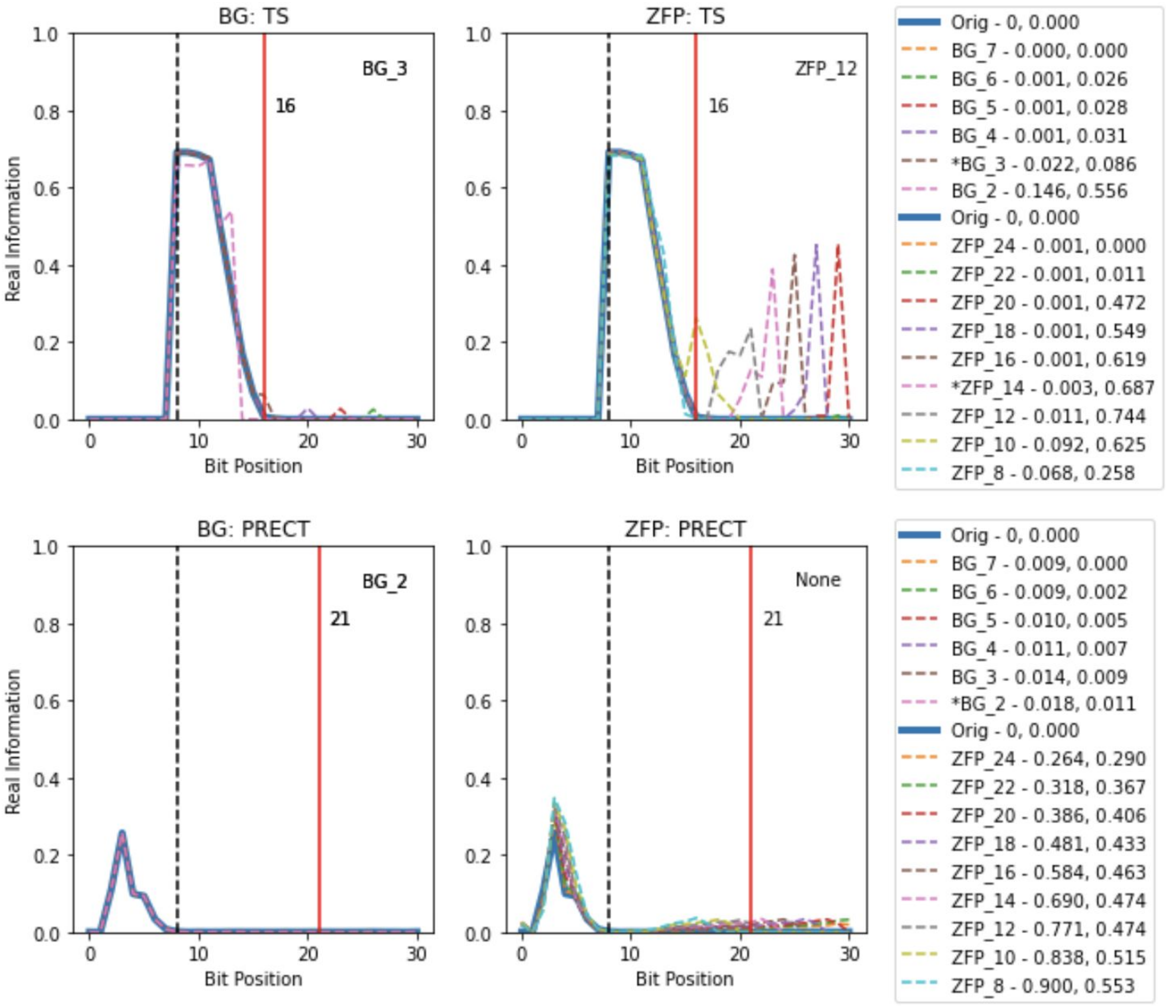
Recap of analysis in DRBSD-22 Paper

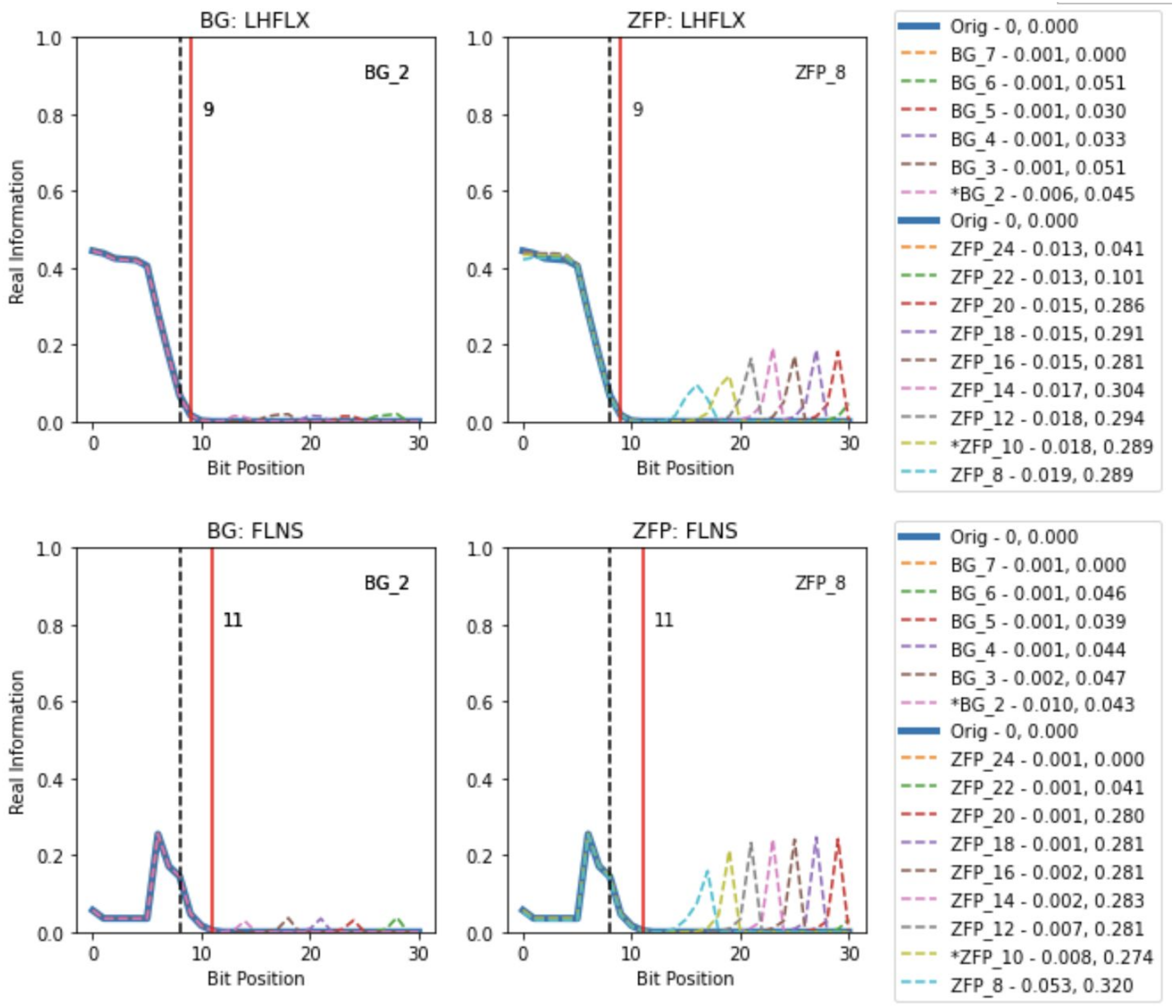
Comparison between real information and DSSIM cutoff parameters to maintain 99% of real information

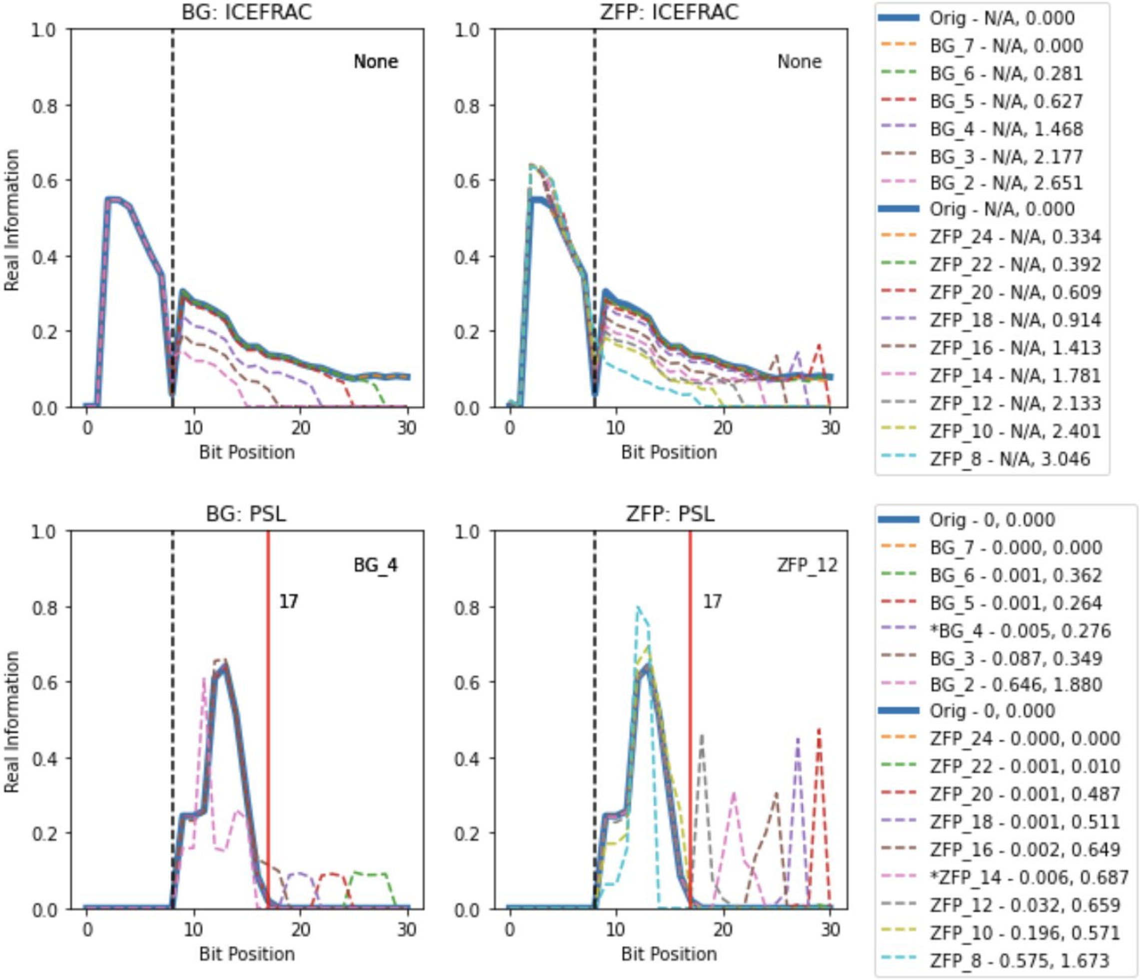
Varable Name	Real Information Content	DSSIM (.95)	(.995)	D (.9995)
FLNS	3	2	3	3
ICEFRAC	N/A	2	4	6
LHFLX	3	2	3	5
PRECT	2	2	3	5
PSL	5	5	6	7
TAUX	2	2	3	4
TS	5	4	5	7
Q200	3	2	4	5
WSPDSRFAV	3	2	3	4
Z500	5	4	5	7

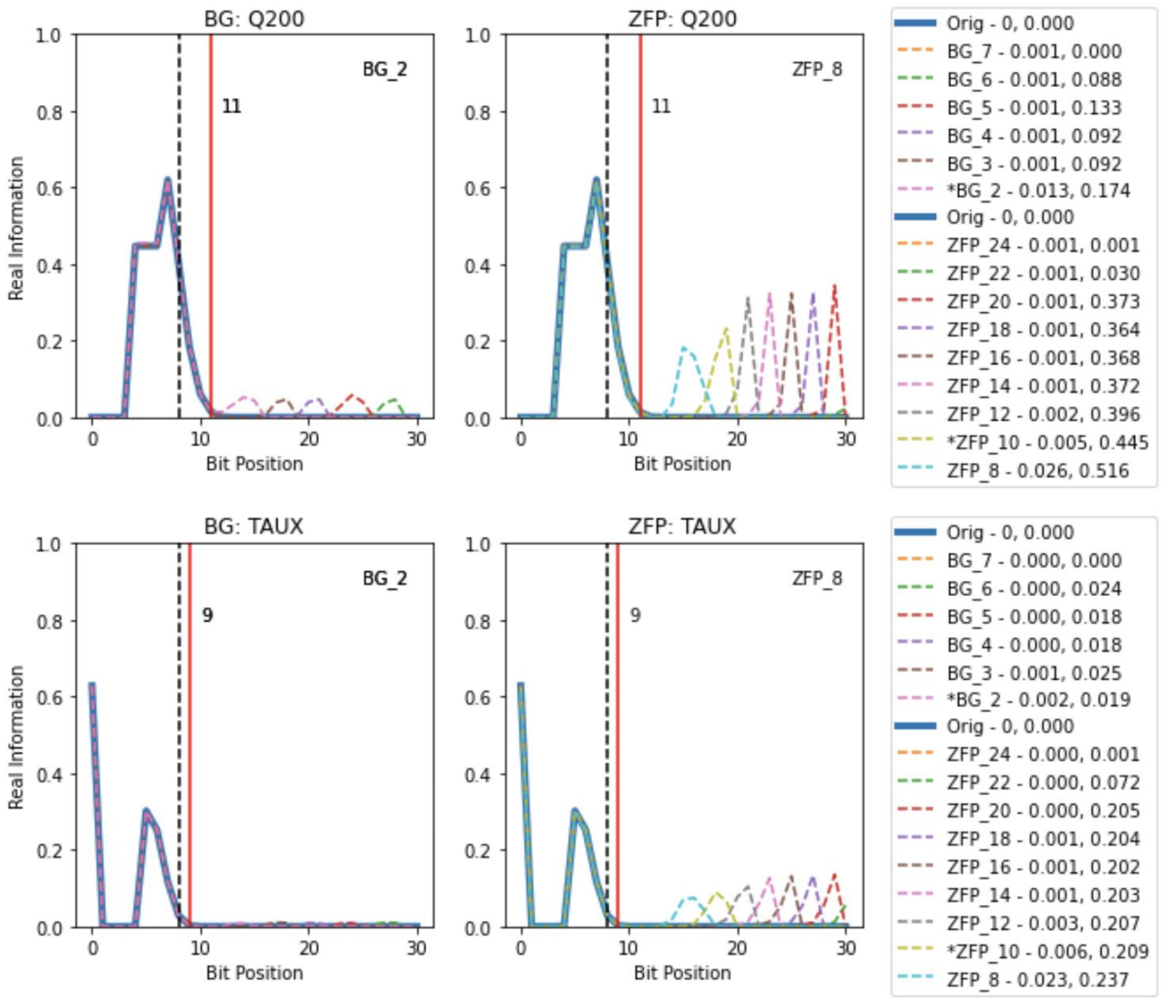
Recap of analysis in DRBSD-22 Paper

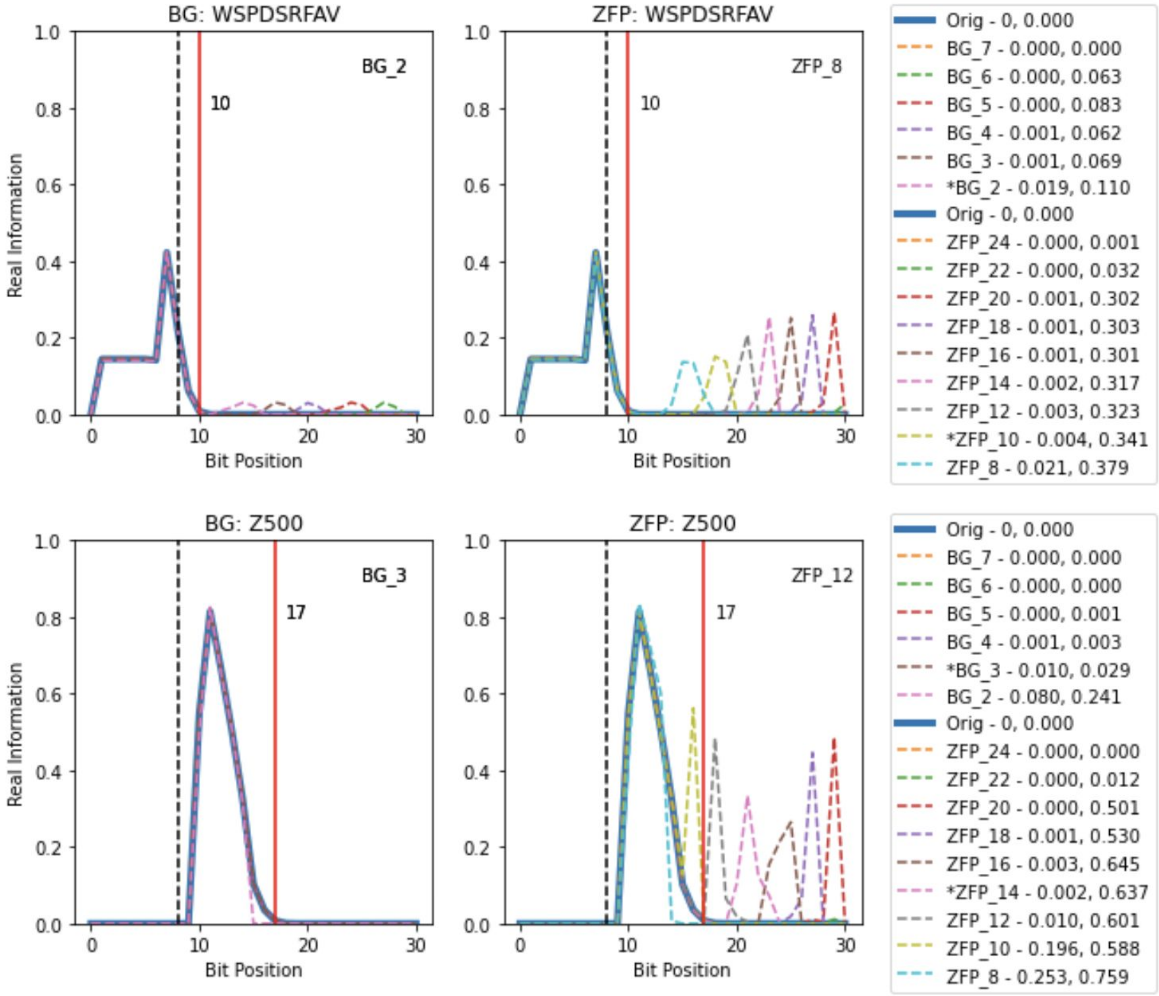
Plots of real information content for variables in original and selected compressed datasets indicating compression artifacts (spikes) and 99% information cutoff bit (vertical red line). Show over next few slides.


BG: TS
BG_3
16
ZFP: TS
ZFP_12
16
Real Information
Bit Position
Orig - 0, 0.000
BG_7 - 0.000, 0.000
BG_6 - 0.001, 0.026
BG_5 - 0.001, 0.028
BG_4 - 0.001, 0.031
*BG_3 - 0.022, 0.086
BG_2 - 0.146, 0.556
Orig - 0, 0.000
ZFP_24 - 0.001, 0.000
ZFP_22 - 0.001, 0.011
ZFP_20 - 0.001, 0.472
ZFP_18 - 0.001, 0.549
ZFP_16 - 0.001, 0.619
*ZFP_14 - 0.003, 0.687
ZFP_12 - 0.011, 0.744
ZFP_10 - 0.092, 0.625
ZFP_8 - 0.068, 0.258
BG: PRECT
BG_2
21
ZFP: PRECT
None
21
Orig - 0, 0.000
BG_7 - 0.009, 0.000
BG_6 - 0.009, 0.002
BG_5 - 0.010, 0.005
BG_4 - 0.011, 0.007
BG_3 - 0.014, 0.009
*BG_2 - 0.018, 0.011
Orig - 0, 0.000
ZFP_24 - 0.264, 0.290
ZFP_22 - 0.318, 0.367
ZFP_20 - 0.386, 0.406
ZFP_18 - 0.481, 0.433
ZFP_16 - 0.584, 0.463
ZFP_14 - 0.690, 0.474
ZFP_12 - 0.771, 0.474
ZFP_10 - 0.838, 0.515
ZFP_8 - 0.900, 0.553


BG: LHFLX
ZFP: LHFLX
BG: FLNS
ZFP: FLNS
Real Information
Bit Position
BG_2
ZFP_8
9
11
Orig - 0, 0.000
BG_7 - 0.001, 0.000
BG_6 - 0.001, 0.051
BG_5 - 0.001, 0.030
BG_4 - 0.001, 0.033
BG_3 - 0.001, 0.051
*BG_2 - 0.006, 0.045
Orig - 0, 0.000
ZFP_24 - 0.013, 0.041
ZFP_22 - 0.013, 0.101
ZFP_20 - 0.015, 0.286
ZFP_18 - 0.015, 0.291
ZFP_16 - 0.015, 0.281
ZFP_14 - 0.017, 0.304
ZFP_12 - 0.018, 0.294
*ZFP_10 - 0.018, 0.289
ZFP_8 - 0.019, 0.289
Orig - 0, 0.000
BG_7 - 0.001, 0.000
BG_6 - 0.001, 0.046
BG_5 - 0.001, 0.039
BG_4 - 0.001, 0.044
BG_3 - 0.002, 0.047
*BG_2 - 0.010, 0.043
Orig - 0, 0.000
ZFP_24 - 0.001, 0.000
ZFP_22 - 0.001, 0.041
ZFP_20 - 0.001, 0.280
ZFP_18 - 0.001, 0.281
ZFP_16 - 0.002, 0.281
ZFP_14 - 0.002, 0.283
ZFP_12 - 0.007, 0.281
*ZFP_10 - 0.008, 0.274
ZFP_8 - 0.053, 0.320


BG: ICEFRAC
ZFP: ICEFRAC
None
None
Real Information
Bit Position
Orig - N/A, 0.000
BG_7 - N/A, 0.000
BG_6 - N/A, 0.281
BG_5 - N/A, 0.627
BG_4 - N/A, 1.468
BG_3 - N/A, 2.177
BG_2 - N/A, 2.651
Orig - N/A, 0.000
ZFP_24 - N/A, 0.334
ZFP_22 - N/A, 0.392
ZFP_20 - N/A, 0.609
ZFP_18 - N/A, 0.914
ZFP_16 - N/A, 1.413
ZFP_14 - N/A, 1.781
ZFP_12 - N/A, 2.133
ZFP_10 - N/A, 2.401
ZFP_8 - N/A, 3.046
BG: PSL
ZFP: PSL
BG_4
ZFP_12
17
17
Real Information
Bit Position
Orig - 0, 0.000
BG_7 - 0.000, 0.000
BG_6 - 0.001, 0.362
BG_5 - 0.001, 0.264
*BG_4 - 0.005, 0.276
BG_3 - 0.087, 0.349
BG_2 - 0.646, 1.880
Orig - 0, 0.000
ZFP_24 - 0.000, 0.000
ZFP_22 - 0.001, 0.010
ZFP_20 - 0.001, 0.487
ZFP_18 - 0.001, 0.511
ZFP_16 - 0.002, 0.649
*ZFP_14 - 0.006, 0.687
ZFP_12 - 0.032, 0.659
ZFP_10 - 0.196, 0.571
ZFP_8 - 0.575, 1.673


BG: Q200
ZFP: Q200
BG_2
ZFP_8
11
11
Real Information
Bit Position
Orig - 0, 0.000
BG_7 - 0.001, 0.000
BG_6 - 0.001, 0.088
BG_5 - 0.001, 0.133
BG_4 - 0.001, 0.092
BG_3 - 0.001, 0.092
*BG_2 - 0.013, 0.174
Orig - 0, 0.000
ZFP_24 - 0.001, 0.001
ZFP_22 - 0.001, 0.030
ZFP_20 - 0.001, 0.373
ZFP_18 - 0.001, 0.364
ZFP_16 - 0.001, 0.368
ZFP_14 - 0.001, 0.372
ZFP_12 - 0.002, 0.396
*ZFP_10 - 0.005, 0.445
ZFP_8 - 0.026, 0.516
BG: TAUX
ZFP: TAUX
BG_2
ZFP_8
9
9
Real Information
Bit Position
Orig - 0, 0.000
BG_7 - 0.000, 0.000
BG_6 - 0.000, 0.024
BG_5 - 0.000, 0.018
BG_4 - 0.000, 0.018
BG_3 - 0.001, 0.025
*BG_2 - 0.002, 0.019
Orig - 0, 0.000
ZFP_24 - 0.000, 0.001
ZFP_22 - 0.000, 0.072
ZFP_20 - 0.000, 0.205
ZFP_18 - 0.001, 0.204
ZFP_16 - 0.001, 0.202
ZFP_14 - 0.001, 0.203
ZFP_12 - 0.003, 0.207
*ZFP_10 - 0.006, 0.209
ZFP_8 - 0.023, 0.237



BG: WSPDSRFAV
ZFP: WSPDSRFAV
BG: Z500
ZFP: Z500
Real Information
Bit Position
BG_2
ZFP_8
BG_3
ZFP_12
10
17
Orig - 0, 0.000
BG_7 - 0.000, 0.000
BG_6 - 0.000, 0.063
BG_5 - 0.000, 0.083
BG_4 - 0.001, 0.062
BG_3 - 0.001, 0.069
*BG_2 - 0.019, 0.110
Orig - 0, 0.000
ZFP_24 - 0.000, 0.001
ZFP_22 - 0.000, 0.032
ZFP_20 - 0.001, 0.302
ZFP_18 - 0.001, 0.303
ZFP_16 - 0.001, 0.301
ZFP_14 - 0.002, 0.317
ZFP_12 - 0.003, 0.323
*ZFP_10 - 0.004, 0.341
ZFP_8 - 0.021, 0.379
Orig - 0, 0.000
BG_7 - 0.000, 0.000
BG_6 - 0.000, 0.000
BG_5 - 0.000, 0.001
BG_4 - 0.001, 0.003
*BG_3 - 0.010, 0.029
BG_2 - 0.080, 0.241
Orig - 0, 0.000
ZFP_24 - 0.000, 0.000
ZFP_22 - 0.000, 0.012
ZFP_20 - 0.000, 0.501
ZFP_18 - 0.001, 0.530
ZFP_16 - 0.003, 0.645
*ZFP_14 - 0.002, 0.637
ZFP_12 - 0.010, 0.601
ZFP_10 - 0.196, 0.588
ZFP_8 - 0.253, 0.759

Repo created for real information research files

Jupyter:

MutualInformation.ipynb

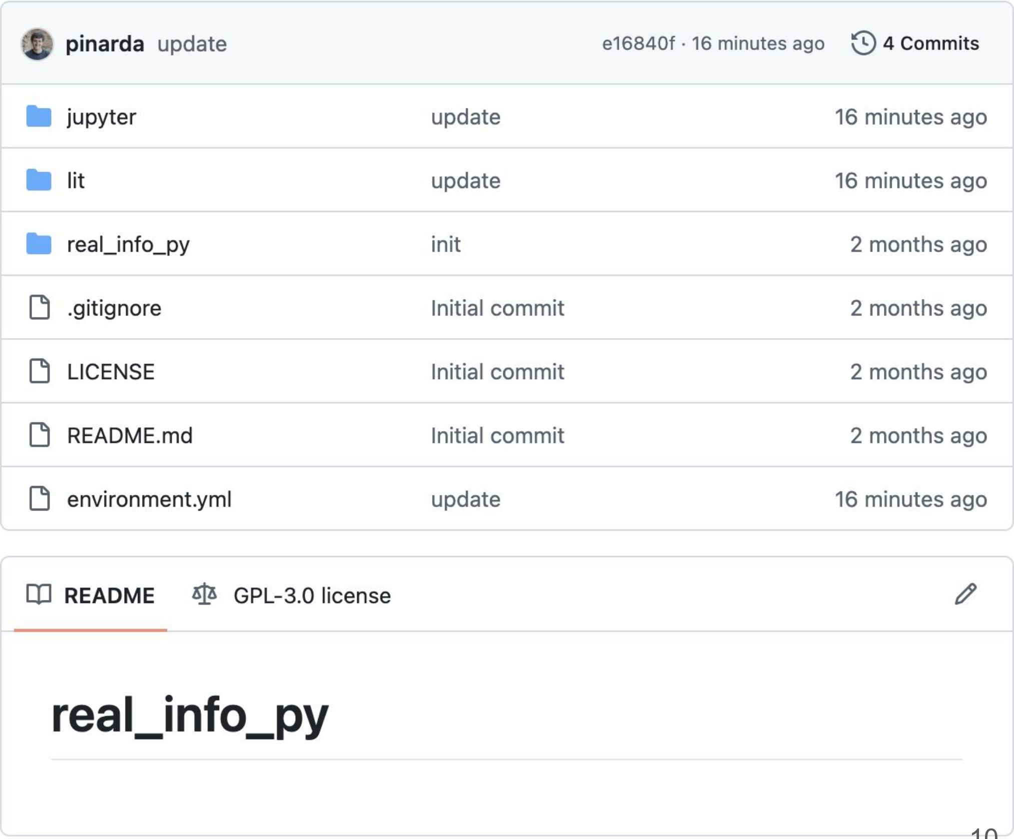
Lit:

DRBSD-22 Paper

Klower Paper

Presentations

Peter Correspondence



Existing Code Functions in ldcpy and jupyter

get_dict_list: Compute current and adjacent bit probabilities

get_mutual_info: Compute mutual info for a set of probabilities

get_real_info: Uses above functions to get real info array

get_bit_cutoff: Get cutoff to maintain given % of real info

get_compression_level: Find level so that, up to cutoff bit, diff between orig and compressed data real info is small

get_real_info_all: Loop over get_real_info for multiple arrays.

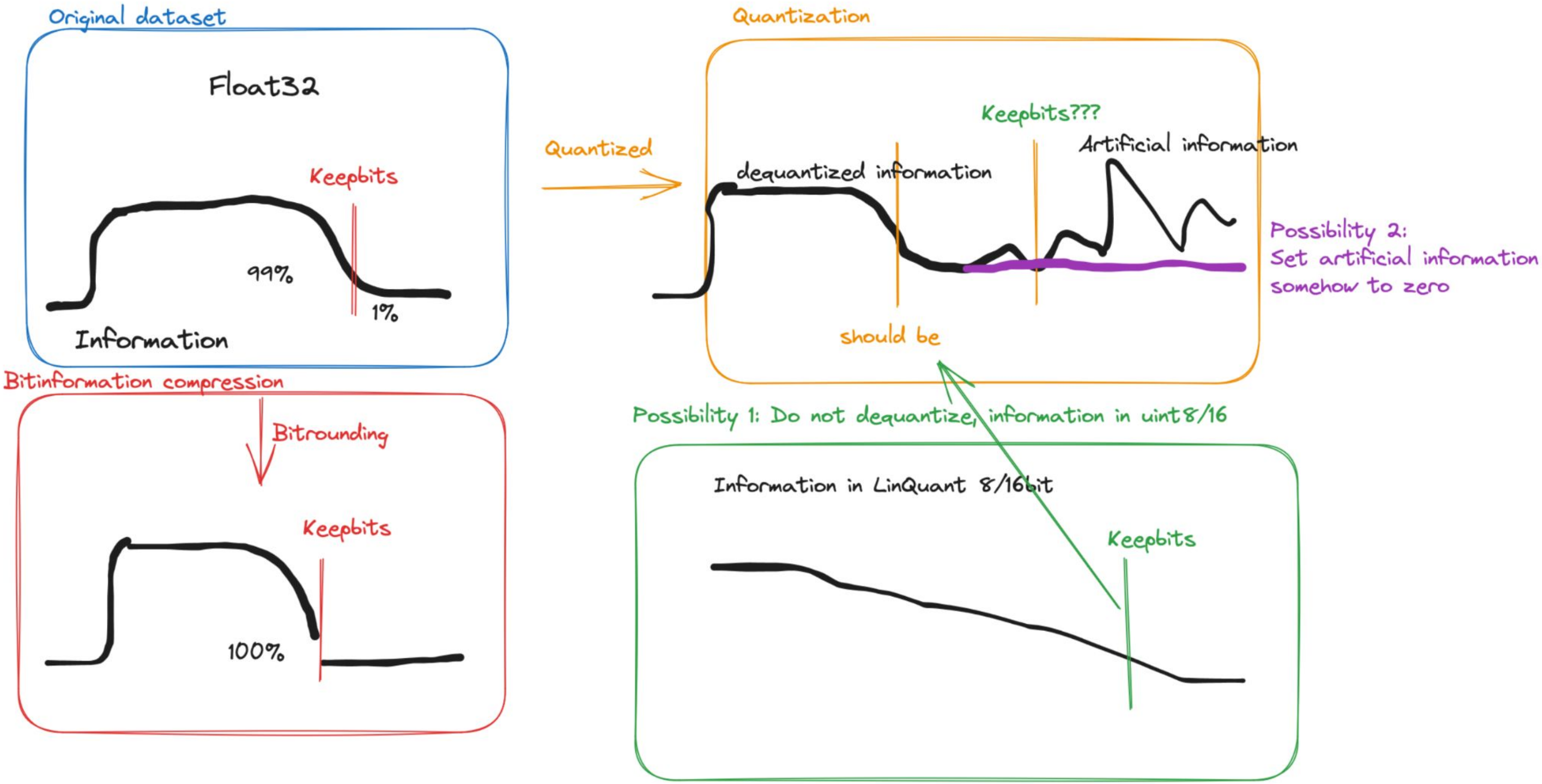
plot_real_info: Create single plot for real info over range of bits.

Future directions

What kinds of artifacts are produced by different compressors?

We have looked at the artifacts introduced by ZFP and BitGrooming, but we can also compress data using SPERR and SZ. Research what is causing these artifacts, if they behave in similar ways across compressors and error modes including the precision and absolute error mode in ZFP (of which we only example the precision mode so far), and can adjustments to the compressors be made to remove these spatial artifacts?

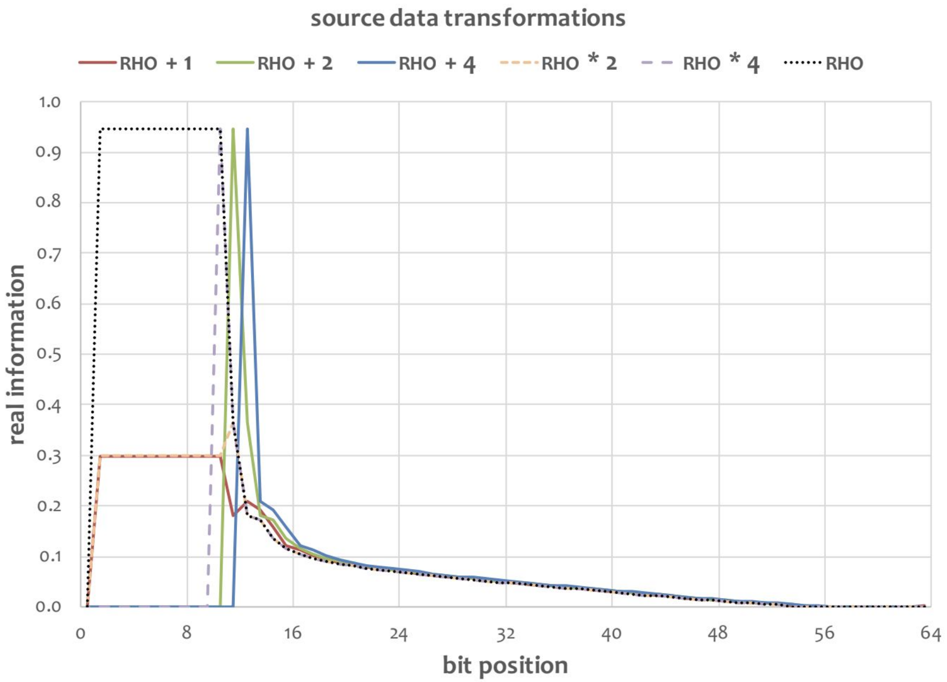
Xbitinfo


Original dataset
Float32
Keepbits
99%
1%
Information
Bitinformation compression
Bitrounding
Keepbits
100%
Quantized
Quantization
Keepbits???
Artificial information
dequantized information
Possibility 2:
Set artificial information
somehow to zero
should be
Possibility 1: Do not dequantize, information in uint8/16
Information in LinQuant 8/16bit
Keepbits

Future directions

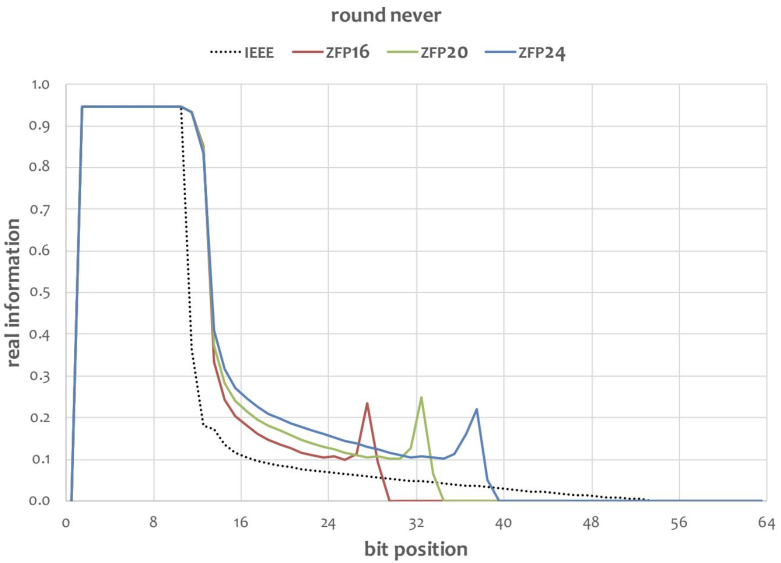
What are the shortcomings of the RIC, when do they apply, and how can the RIC be improved?

The same numerical values stored in different ways, such as with exponent values that differ from neighboring points, causes alterations in the computed RIC. This will affect data that spans a large range of exponent values in a small spatial area, such as PRECT (precipitation rate) data. Adjustments to the process of computing RIC can be made to account for these shortcomings. Additionally, the RIC also operates 1-dimensionally, where the dataset is flattened and neighboring bits are considered to be only the bits adjacent in the resulting bit stream. The measure could be extended into two dimensions to more accurately capture the notion of spatial structure.



-1 => 01111111110 (23.3%)
0 => 01111111111 (30.1%)
1 => 10000000000 (46.6%)

0 => 01111111111 (23.3%)
1 => 10000000000 (30.1%)
2 => 10000000001 (46.6%)



Future directions

How can the cutoff bit recommended by the RIC be translated into a parameter setting such as relative error for a given compression algorithm?

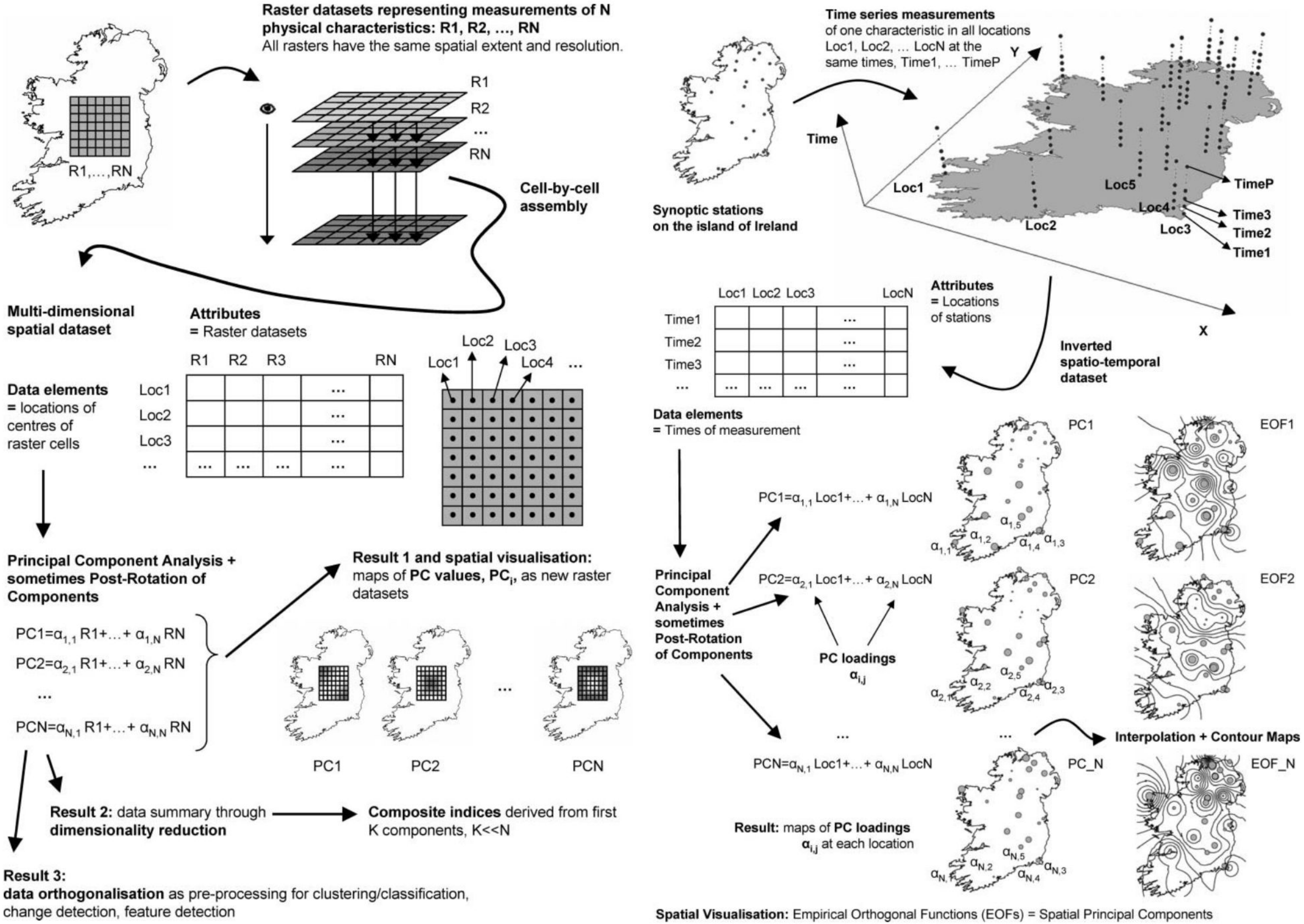
We will explore the parameter settings that align with the RIC cutoff bit and the corresponding levels of compression that can be obtained. This can be done via brute-force, where multiple compression algorithms and settings are applied, and the parameters that lead to median errors near the location of the bit cutoff line are a ``match" for that RIC cutoff.

Additionally, explore the connection between the RIC and other quantities of interest such as the DSSIM.

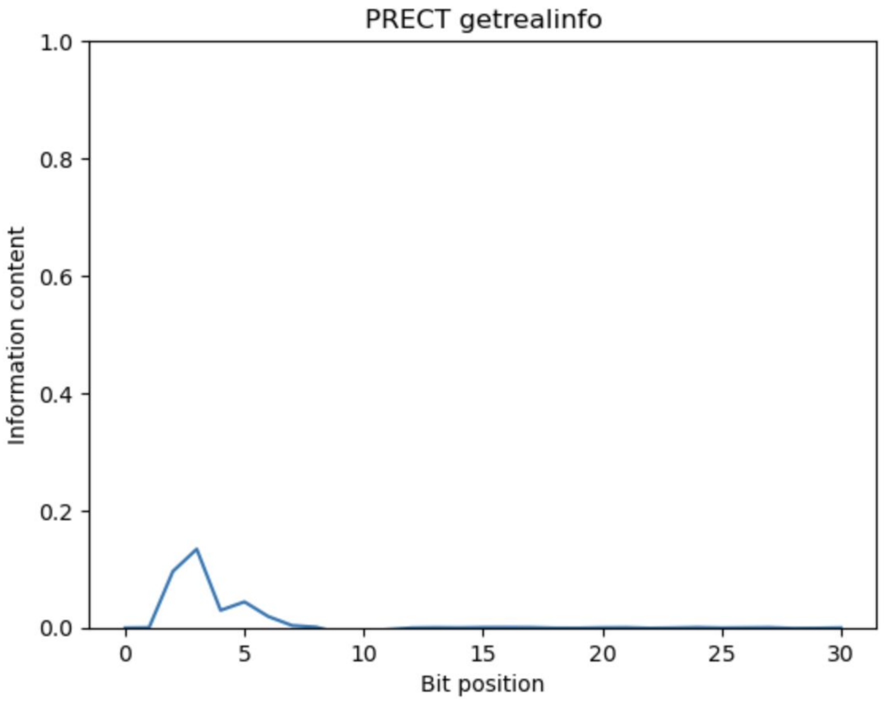
END

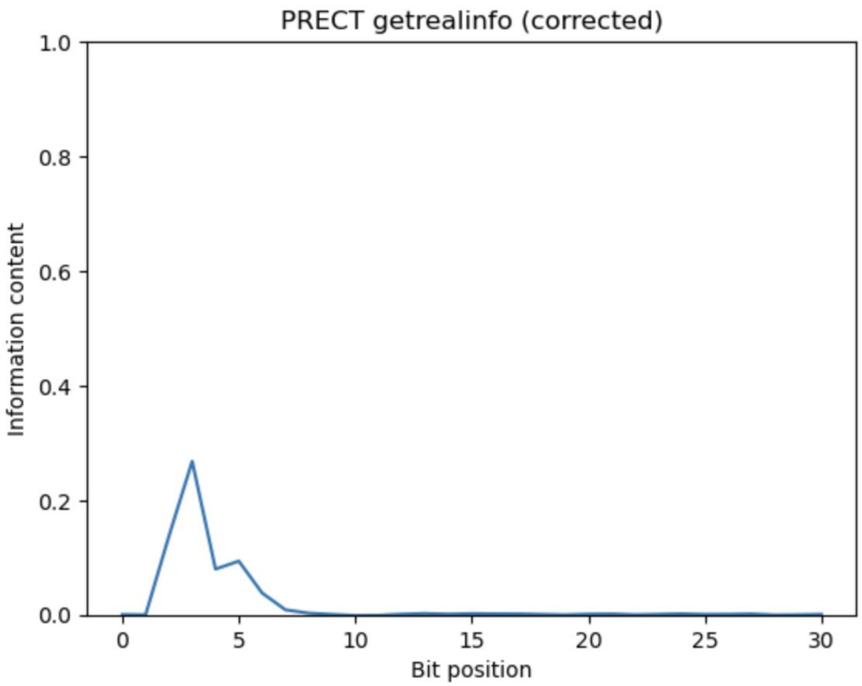
Other Ideas

- Extensions to the Real information - look at additional neighboring values
- Fitting spatial model and making comparisons to range parameter/error
- Connections to PCA?


Raster datasets representing measurements of N physical characteristics: R1, R2, …, RN
All rasters have the same spatial extent and resolution.
R1,…,RN
R1
R2
…
RN
Cell-by-cell assembly
Multi-dimensional spatial dataset
Attributes
= Raster datasets
R1 R2 R3 RN
Loc1 Loc2 Loc3 …
Loc1 Loc2 Loc3 Loc4 …
Data elements
= locations of centres of raster cells
Principal Component Analysis + sometimes Post-Rotation of Components
PC1=α1,1 R1+…+ α1,N RN
PC2=α2,1 R1+…+ α2,N RN
…
PCN=αN,1 R1+…+ αN,N RN
Result 1 and spatial visualisation: maps of PC values, PCi, as new raster datasets
PC1 PC2 … PCN
Result 2: data summary through dimensionality reduction
Composite indices derived from first K components, K<<N
Result 3:
data orthogonalisation as pre-processing for clustering/classification, change detection, feature detection
Synoptic stations on the island of Ireland
Time series measurements of one characteristic in all locations Loc1, Loc2, … LocN at the same times, Time1, … TimeP
Time
Y
X
Loc1 Loc2 Loc3 Loc4 Loc5
TimeP Time3 Time2 Time1
Attributes
= Locations of stations
Loc1 Loc2 Loc3 LocN
Time1 Time2 Time3 …
Inverted spatio-temporal dataset
Data elements
= Times of measurement
Principal Component Analysis + sometimes Post-Rotation of Components
PC1=α1,1 Loc1+…+ α1,N LocN
PC2=α2,1 Loc1+…+ α2,N LocN
…
PCN=αN,1 Loc1+…+ αN,N LocN
PC loadings αi,j
PC1 EOF1
PC2 EOF2
PC_N EOF_N
Interpolation + Contour Maps
Result: maps of PC loadings αi,j at each location
Spatial Visualisation: Empirical Orthogonal Functions (EOFs) = Spatial Principal Components

Corrected real information jupyter notebook





Corrected real information jupyter notebook