Capturing the Intangible Concept of Information In Hydrological Simulation
——Perspectives From Shannon and Kolmogorov

Pan Baoxiang　Cong Zhentao

Institute of Hydrology and Water Resources
Tsinghua University

June 4, 2015

Outline

- Introduction
- Shannon Information & Kolmogorov Complexity
- Case Study
- Discussion & Conclusion

Introduction

Questions:

- Given the p.d.f. of the rainfall amount of a certain year, through how many guesses could we reach its interval estimation to a fixed accuracy?
- Give an efficient description of an annual rainfall series.

Introduction

Questions:

- Given the p.d.f. of the rainfall amount of a certain year, through how many guesses could we reach its interval estimation to a fixed accuracy?
- Give an efficient description of an annual rainfall series.

Information is Bits + Context.

Measuring Information Contents

	Shannon Entropy	Kolmogorov Complexity
Definition	$H(X) = -\Sigma p(x) log p(x)$ $h(X) = -\int f(x) log f(x) dx$	The length (in bits) of the shortest computer program that prints the sequence and then halts.
Focus	Random Source Object irrelevant	Object Probability irrelevant
Property	$H(X, Y) \leq H(X) + H(Y)$	Uncomputability
Estimator	p.d.f Estimator	Compressor
Dimension	Bit	

Measuring Information Connections

	Mutual Information(S)	Mutual Information(K)
Definition	$I(X;Y) = \sum p(x,y) log \frac{p(x,y)}{p(x)p(y)}$ $I(X;Y) = \int f(x,y) log \frac{f(x,y)}{f(x)f(y)} dxdy$	$KC(X) + KC(Y) - KC(X,Y)$
Focus	Random Source Object irrelevant	Object Probability irrelevant
Property	Symmetry	Uncomputability
Estimator	KNN+SVR	$ZIP(X) + ZIP(Y) - ZIP(X,Y)$
Dimension	Bit	

Bits in Hydrological Simulation Context

$$\text{Hydrological Simulation}\begin{cases} \text{observation}\begin{cases} Input\ \ Measurements\ X_i \\ Output\ \ Measurements\ X_o \end{cases} \\ \\ \text{Model}\qquad\begin{cases} Output\ \ Simulation\ X_s \\ State\ \ Variable\ \ Simulation\ S \\ Parameters \\ ModelStructure \end{cases} \end{cases}$$

Bits in Hydrological Simulation Context

$$
\textit{Hydrological Simulation}
\begin{cases}
\textit{observation} \begin{cases} \textit{Input Measurements } X_i \\ \textit{Output Measurements } X_o \end{cases} \\[2em]
\textit{Model} \quad \begin{cases} \textit{Output Simulation } X_s \\ \textit{State Variable Simulation } S \\ \textit{Parameters} \\ \textit{ModelStructure} \end{cases}
\end{cases}
$$

Bits in Hydrological Simulation Context

- For a time/frequency/feature-space domain base,
  - X_i,X_o,X_s,S are represented by their coordinates,
    - Entropy / Kolmogorov Complexity of these coordinates represents the complexity of expressing the signals with that base:
      $H(X_o)$ / $KC(X_o)$: Bits are required to depict X_o.
    - Shannon / Kolmogorov Mutual Information between coordinates represents the information connection of the signals at that base:
      $MI(X_i, X_o)$: Bits provided by X_i.
      $MI(X_s, X_o)$: Bits provided by Model.

Bits in Hydrological Simulation Context

- For a time/frequency/feature-space domain base,
  - X_i,X_o,X_s,S are represented by their coordinates,
    - Entropy / Kolmogorov Complexity of these coordinates represents the complexity of expressing the signals with that base:
      $H(X_o)$ / $KC(X_o)$: Bits are required to depict X_o.
    - Shannon / Kolmogorov Mutual Information between coordinates represents the information connection of the signals at that base:
      $MI(X_i, X_o)$: Bits provided by X_i.
      $MI(X_s, X_o)$: Bits provided by Model.

$$Aleatory\ Uncertainty = H(X_o)/KC(X_o) - MI(X_i, X_o)$$

$$Epistemic\ Uncertainty = MI(X_i, X_o) - MI(X_s, X_o)$$

Bits in Hydrological Simulation Context

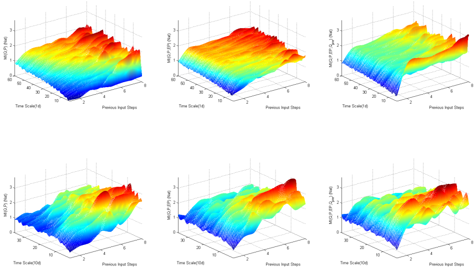
- For a time/frequency/feature-space domain base,
  - X_i,X_o,X_s,S are represented by their coordinates,
    - Entropy / Kolmogorov Complexity of these coordinates represents the complexity of expressing the signals with that base:
      $H(X_o)$ / $KC(X_o)$: Bits are required to depict X_o.
    - Shannon / Kolmogorov Mutual Information between coordinates represents the information connection of the signals at that base:
      $MI(X_i, X_o)$: Bits provided by X_i.
      $MI(X_s, X_o)$: Bits provided by Model.

$$Aleatory\ Uncertainty = H(X_o)/KC(X_o) - MI(X_i, X_o)\ \textit{Noise}$$

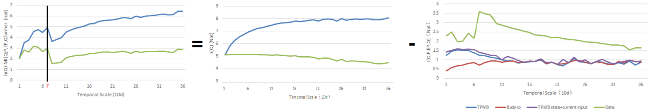
$$Epistemic\ Uncertainty = MI(X_i, X_o) - MI(X_s, X_o)\ \textit{Signal}$$

Case Study 1: How water-heat correlation emerges through temporal upscaling

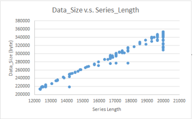
Relation of MI,Scale,Previous-Input-Step



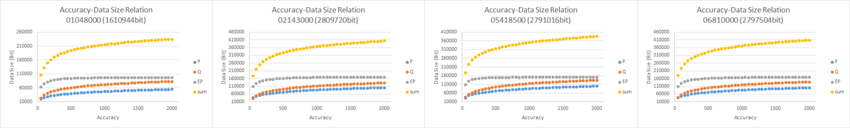
Signal-to-noise Ratio Across Temporal Scales



Case Study 2: To what ratio can hydrological data be compressed



Huffman Encoding —— Lisp Implementation



Case Study 2: To what ratio can hydrological data be compressed

Data set	Constant	Linear	Uniform white	Gaussian white	Sin 1	Sin 100	Leaf Q	Leaf P
file size	50 000	50 000	50 000	50 000	50 000	50 000	14 610	14 610
$\frac{H}{\log N}$	0.0	99.9	99.9	86.3	96.0	92.7	42.1	31.0
SNR	NaN	255.0	255.6	108.0	307.4	317.8	42.6	39.9
Uncompressed formats								
BMP	102.2	102.2	102.2	102.2	102.2	102.2	407.4	407.4
WAV	100.1	100.1	*100.1*	100.1	100.1	100.1	100.3	100.3
HDF_NONE	100.7	100.7	100.7	100.7	100.7	100.7	102.3	102.3
Lossless compression algorithms								
JPG_LS	12.6	12.8	110.6	94.7	12.9	33.3	33.7	49.9
HDF_RLE	2.3	2.7	101.5	101.5	3.2	92.3	202.3	202.3
WAVPACK	*0.2*	1.9	103.0	*87.5*	2.9	25.6	38.0	66.2
ARJ	0.3	1.0	100.3	88.0	3.1	1.9	33.7	40.0
PPMD	0.3	2.1	102.4	89.7	3.6	1.4	*27.7*	*36.4*
LZMA	0.4	0.9	101.6	88.1	1.9	1.2	31.0	37.8
BZIP2	0.3	1.8	100.7	90.7	3.0	2.3	29.8	40.5
PNG	0.3	*0.8*	100.4	93.5	*1.5*	*0.8*	40.2	50.0
GIF	2.3	15.7	138.9	124.5	17.3	32.0	38.8	45.9
TIFF	2.0	2.4	101.2	101.2	2.9	91.2	201.5	201.5

Weijs S V, Giesen N, Parlange M B. Data compression to define information content of hydrological time series[J]. Hydrology and Earth System Sciences, 2013, 17(8): 3171-3187.

Discussion & Conclusion

Coding is not merely an E.E. trick. It's hard to tell where data science stops and coding starts.

- Significance Digging
- Application Expansion

https://github.com/morepenn

END