Information Content of Ocean Colour Data

Dr. A. Neumann, H. Krawczyk
German Aerospace Centre DLR
Remote Sensing Technology Institute
Marine Remote Sensing

- What is Information Content?
- Instrument and Target Aspects
- Analysis of Information Content
- Retrievability of Parameters
- Multivariate Analysis - Principal Components and Intrinsic Dimensionality

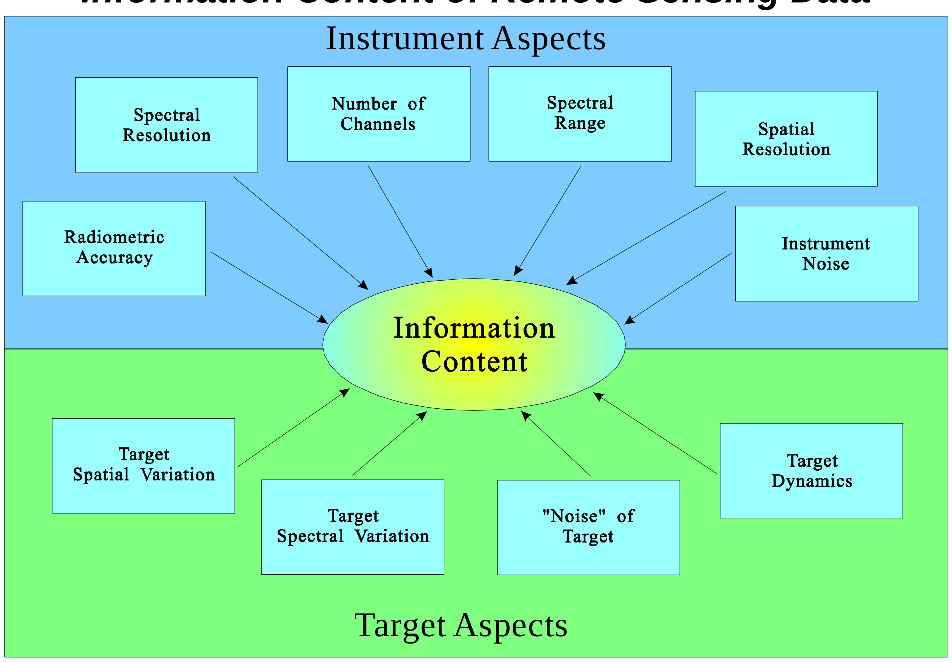
What is Information Content?

- ... the attempt to measure (i.e. quantify, enumerate) the “valuable information” in a set or stream of data
- different approaches worked out in information theory, statistics and applied sciences, depending on the problem to be solved (optimal coding strategies, data compression, parameter retrieval, ...)
- in the context of ocean colour remote sensing: requirements to spectral characteristics (resolution, number of channels, SNR) from the interpretation/applicational perspective

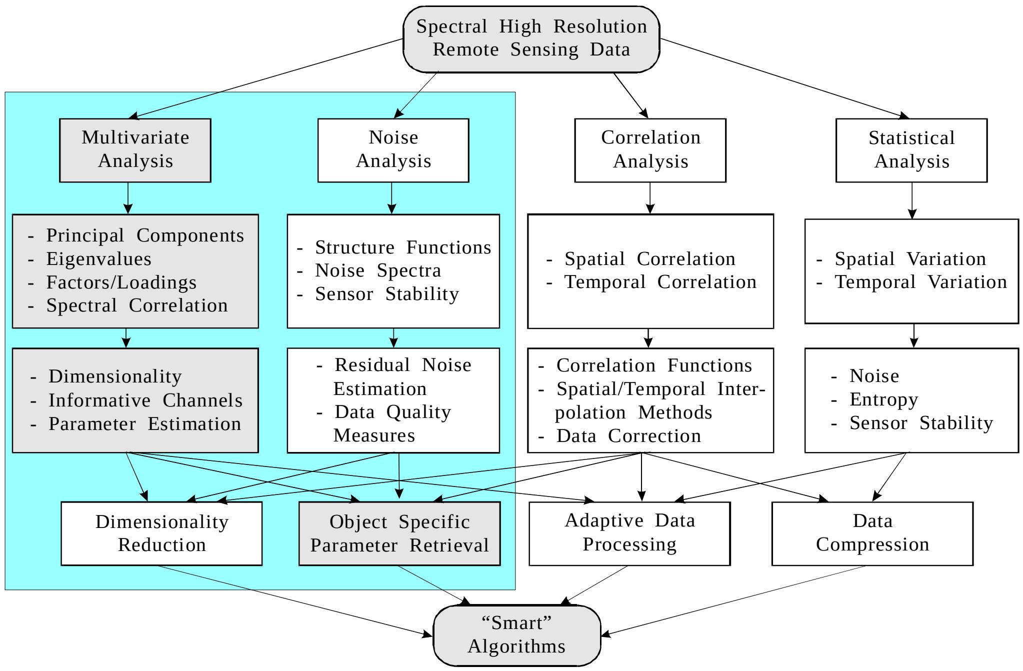
Parameters to Quantify or Handle Information Content:

- entropy (statistics, data coding)
- correlation (spectral, spatial)
- dominant factors (Factor Analysis)
- significant Ortho-Components (Principal Component Analysis, Karhunen-Loewe Transformation)
- *in our context leading to*
  - ⇨ ***intrinsic dimensionality analysis***
  - ⇨ ***"retrievability" of parameters***
  - ⇨ ***error analysis of retrieval algorithms***

Information Content of Remote Sensing Data

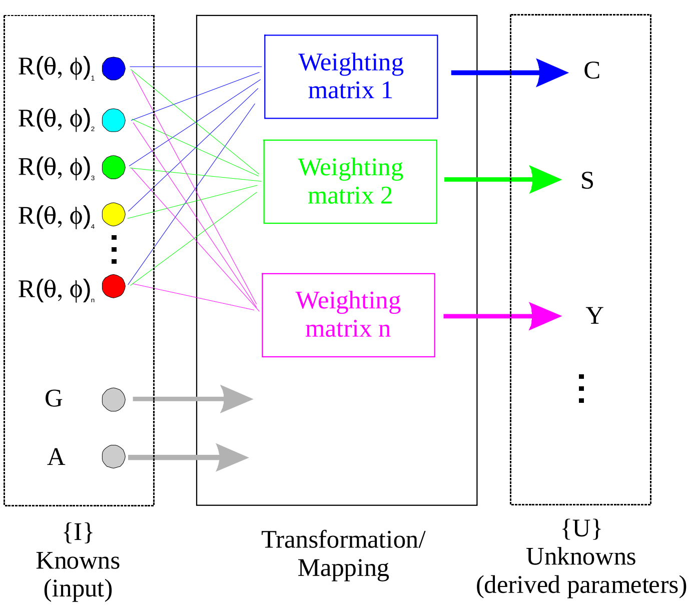


Analysis of Information Content



Training Course on Remote Sensing of Ocean Colour
Ahmedabad, India, February 2001
Information Content of Ocean Colour Data
andreas.neumann@dlr.de

The Remote Sensing Problem



! The inverse task is analytically not solvable

? Which and how many parameters can be retrieved

? Which and how many channels contribute how much to each parameter

? How to construct an „optimal“ algorithm

How to Perform an Information Analysis?

- Using experimental data, i.e. satellite and in-situ measurements:
  - quality/accuracy of the data often not known
  - influencing physical parameters not under control
  - problem: large number (>10^4) of data sets for various combinations of parameters necessary

- Using bio-optical and radiative transfer models:
  - clear understanding of mathematical and physical interdependencies
  - variability and number of parameters under control
  - problem: adequateness of the model for practical application

Retrievability of Parameters - Basics

- We have measurements/simulations of spectral signatures, i.e. values of radiance/reflectance at certain λ (channels) with certain $\Delta\lambda$ (spectral halfwidth) and a given SNR and corresponding (bio-) physical parameters to be retrieved from the data $\{L_1, ..., L_N, C, S, Y, \tau_A, ...\}$

- generally no link between ONE parameter and ONE channel
- usually significant spectral correlation, at least in spectral subranges
- influence of the parameters spectrally overlapping
- statistical behaviour and interdependencies of parameters not predictable a priori

⇨ first (and simple) test: numerical analysis
⇨ second investigation: multivariate analysis

Spectral Overlap of Different Constituents

Spectral range	Chlorophyll	Gelbstoff	Sediment
blue	absorption	strong absorption	scattering
green	scattering	weak absorption	scattering
red	scattering, absorption fluorescence	absorption	scattering

Atmosphere/aerosol: scattering at all wavelengths

Retrievability of Parameters - Numerical Analysis

- Generate a large number of M spectra $\{L_1, ..., L_N\} = f(C, S, Y, \tau_A\}$ using radiative transfer model, N - number of available spectral channels
- degrade the “ideal” values by a given NEΔL or relative “measurement” error
- anlyse, how many spectra can be distinguished, i.e. are at least in one channel different to each other
- example for MOS channels: data set of 140.000 spectra for independently varying parameters of C = 0..100 µg/l, a_Y(440nm) = 0..1, b_S(550nm)=0..10, τ_A = 0..0.25

Relative Error [%]	0	1	3	5	7	10
Discriminated Spectra [%]	97	69	25	12	8	4

Numerical Analysis - Conclusion

- for theoretical case of zero error the result is not too bad, disillusioning for realistic error values
- underlines high requirements to measurement accuracy
- more detailed analysis of error propagation and influence of single parameters possible, however, this seems not the adequate tool to handle correlations and interdependencies

⇨ ill-defined, multivariate problem which needs more sophisticated tools for analysis

Retrievability of Parameters - Multivariate Analysis

- Multivariate analysis (e.g. factor analysis, principal component analysis) provide an effective mathematical tool to analyse large data sets in the presence of correlation and interdependencies
- the principle is an “optimal” transformation of the original data set to an orthogonal system containing identical information, accounting for (spectral) correlation
- leads to diagonalisation of the covariance matrix
- is a purely mathematically defined tool, not accounting for physical background represented by the data

Principal Component Analysis (PCA)

- let Cov{$L_j/\Delta L_j$}, j=1..N, be the error normalised spectral covariance matrix of the data set, the principal components are computed as:

$$PC_k = \sum_{m=1}^{N} \frac{U_{km}(L_m - \overline{L}_m)}{\Delta L_m}$$

- where:
  - PC_k - kth principal component, k from 1 to N
  - N - number of spectral bands used in the original data set
  - $\underline{U}_k$ - kth eigenvector of the covariance matrix
  - L_m - radiance in the mth spectral band
  - $\overline{L}_m$ - mean radiance value in the mth spectral band
  - ΔL_m - noise-equivalent radiance, measurement error.
- Due to the transformation principle (diagonalisation of the covariance matrix) the PC are uncorrelated (orthogonal) and contain successively maximal information in the sense of signal variance:

◷PC_i◷ > ◷PC_j◷ for i>j

PCA - SNR and Instrinsic Dimensionality

- Since the computation was done using error-normalised radiances, the eigenvalue gives directly the SNR of the corresponding PC:

$$\lambda_k = SNR^2(PC_k)$$

  with λ_k beeing the kth eigenvalue of Cov{$L_j/\Delta L_j$}

- based on this we can define the ***intrinsic dimensionality*** of the analysed data set as:

$$D = \max(k) \text{ with } \sqrt{\lambda_k} >> 1, D<N.$$

- The intrinsic dimensionality gives a measure, how many PCs contain significant information (in the sense of variance), i.e. are necessary to reconstruct the full information content of the original data set. It is , however, an unsharp definition.

What does Instrinsic Dimensionality Tell Us?

- It tells us how many ***statistically independent parameters*** can be retrieved from the data set in principle (but not which ones and how!)
- it tells us whether the inverse task is solvable in general (D ♦ number of desired parameters)
- it ***does not say***, that we need only D spectral channels, because:
  each parameter retrieval from the spectral measurement can be seen as a statistical estimate ⇨ general rule: the larger the number of measurements, the higher the achievable accuracy!

- examples:
  – for simulated MOS-data (varying C, S, Y, τ_A, α) we get D up to 6
  – real MOS-data over water D ~ 5

IOCCG International Ocean Colour Coordinating Group
Training Course on Remote Sensing of Ocean Colour
Ahmedabad, India, February 2001
Information Content of Ocean Colour Data
DLR
andreas.neumann@dlr.de

Appendix: Technical Parameters MOS-IRS

Parameter	**MOS-A**	**MOS-B**	**MOS-C**
Spectral Range [nm]	755 - 768	408 - 1010	SWIR
No. of Channels	4	13	1
Wavelengths [nm]	756.7; 760.6; 763.5; 766.4 O_2A-band	408; 443; 485; 520; 570; 615; 650; 685; 750; 870; 1010 815; 945 (H_2O- vapor)	1600
spectral halfwidth [nm]	1.4	10	100
Field of View **along track x [deg]** **across track [deg]**	 0.344 13.6	 0.094 14.0	 0.14 13.4
Swath Width [km]	195	200	192
No. of Pixels	140	384	299
Pixel Size x*y [km^2]	1.57x1.4	0.52x0.52	0.52x0.64
Measuring Range **Lmin [μWcm^{-2}nm^{-1}sr^{-1}]** **Lmax [μWcm^{-2}nm^{-1}sr^{-1}]**	 0.1 40	 0.2 65	 0.5 18