Compressing Massive Geophysical Data Sets Using Vector Quantization

Amy BRAVERMAN*
California Institute of Technology
Mail Stop 169-237, Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, CA 91109-8099 USA
Tel: 0011-1-818-354-6168
Fax: 0011-1-818-393-4619
email: amy@jord.jpl.nasa.gov

Topics: D7, D6, D4. Presentation preference: oral.
*Corresponding Author

Massive geophysical data sets like those obtained from the Multi-angle Imaging SpectroRadiometer (MISR) and the International Satellite Cloud Climatology Project (ISCCP) are becoming commonplace, but data volume makes their analyses difficult. Traditional strategies such as sub-setting, sub-sampling, and aggregating to lower spatio-temporal resolution sacrifice global coverage, ignore data, and can aggregate away high-resolution, multivariate relationships. The method proposed here reduces data volume while approximately preserving high-resolution data structure. It creates a set of faithful statistical summaries that serve as a proxy for the original data. Researchers can use these to conduct global, exploratory analyses, and identify phenomena of interest. These can then be further investigated using the original, high-resolution data.

The procedure uses an extension of a lossy data compression algorithm, Entropy-constrained Vector Quantization (ECVQ; Chou, Lookabaugh, and Gray, 1989, Braverman, 2000), to summarize geophysical data on a one degree by one degree, monthly global grid. Each cell defines a subset of N data points in C-dimensional space, where C is the number of geophysical parameters. The N data points are partitioned into groups, and each group represented by its C-dimensional mean vector. The set of mean vectors and numbers of data points they represent are a summary (or equivalently, a compressed version) of the original cell data. ECVQ forms groups in a way that optimally trades-off information loss suffered using the summary in place of the full cell data set against data reduction achieved. By applying the algorithm in a coordinated way to all cells, a collection of summaries is obtained that reflect high-dimensional data structure both within and between cell data sets. The statistical mean squared errors of these summaries as estimators of their parent data are also reported so one can gauge how well functions of the original data are estimated by corresponding functions of their compressed counterparts.

This method was developed for purposes of creating global (Level 3) MISR data products, and has also been applied to ISCCP (DX) data. Examples and computational issues drawn from both MISR and ISCCP are discussed.