

COMPRESSION AND ANALYSIS OF VERY LARGE IMAGERY DATA SETS USING SPATIAL STATISTICS

James A Shine,
George Mason University, and
US Army Topographic Engineering Center
7701 Telegraph Road
Alexandria VA 22315-3864
jshine1@gmu.edu

ABSTRACT:

As remote sensing instruments evolve, the size of imagery data sets derived from remote sensing continues to increase. Several satellites currently offer resolution of 1 meter per pixel or better. At this resolution, even a small geographic area leads to a very large data set; 1 square mile, for example, is represented by approximately 2.6×10^6 pixels. Many sensors are now multispectral or even hyperspectral, increasing the size of the data set by up to 10^2 . Processing images for classification or mapping purposes thus poses an increasing computational challenge.

This paper describes the use of spatial statistics to compress the size of large 1-meter imagery data sets. The images were taken over locations in the United States using a CAMIS (Computerized Airborne Multispectral Imaging System) instrument flown in an airplane and registered by trained image analysts. Models of spatial variation are first computed on an entire image, then on subsampled sets of the image. Parameters of the models are used to compress the original image. In some cases it is possible to compress data several orders of magnitude without substantially degrading results of subsequent analysis.

INTRODUCTION:

An ongoing challenge for the image analysis community is the increasing size and complexity of imagery data sets. Pixel resolution continues to improve; many remote sensors now have a pixel resolution of 1 meter or less. In addition, images are often multivariate, with multispectral and hyperspectral sensors becoming more prevalent. LANDSAT TM imagery contains 7 bands; hyperspectral imagery such as AVIRIS, HYDICE, and SEBASS contain over 200 bands. More data is being collected over time as well by such sensors as those in the Earth Observing program (1 terabyte of data a day). All of these factors increase the size of images to be analyzed for classification, terrain analysis and other purposes. As a result, compression of these image sets becomes an important issue.

Spatial statistics offers a new approach to compression of very large imagery data sets. Patterns of spatial correlation have been used to map and model imagery data [1,2]. These patterns can also be used to quantify the degree to which an image can be compressed without losing important information. Of particular interest to those applying spatial statistics to imagery data is the degree to which an image can be compressed without losing its spatial information. This is the focus of the experiments and results reported here.

METHODOLOGY:

Some data observations have spatial locations associated with them. Early heuristic approaches provided empirical evidence that spatial information could improve the estimation of data values at new locations based on the data values at existing

locations. Later, Matheron developed the theory of regionalized random variables, which provided a theoretical underpinning to these heuristic approaches. [3]

The two main assumptions of regionalized random variable theory are stationarity and isotropy. Under stationarity, the variance between points does depend on the distance, as does the correlation, but at a given distance h these two values should be the same regardless of location. The mean is also the same regardless of location. Isotropy assumes that data statistics are independent of direction; the mean, variance and correlation between points does not depend on whether the points run north-south, east-west, or some other direction. Since these conditions are often not met with real data, assumption relaxations such as the Intrinsic Hypothesis (where the variance may be unbounded) and quasi-stationarity (where stationarity applies in a neighborhood of the data, but not in the entire data domain) have been developed.

The variation in spatial data can be divided into two components. Stochastic variation is ordinary variation such as occurs in nonspatial data, and spatial variation is variation that depends on the distance between points.

The first step in spatial data analysis is the computation and plotting of a variogram. All the observations in a data set are compared to all other observations, and for each distance h between points, a semivariance function γ which is half of a traditional variance is computed. The stochastic variation referred to in the introduction shows up as a baseline or "nugget" variance which is the value of γ at $h=0$. The spatial variation shows up in the values of γ as h increases. A typical image variogram is shown in Figure 1.

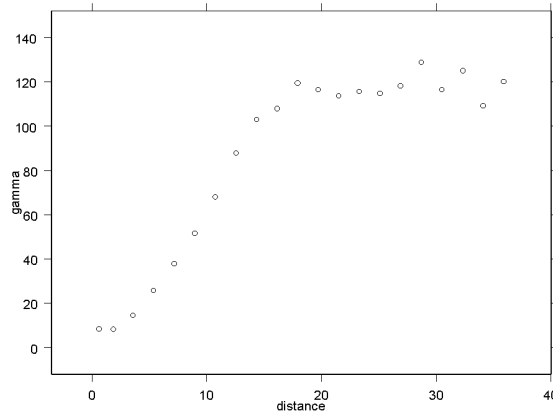


Figure 1: A typical image variogram

Variograms can be used for several purposes important to imagery analysis: the scale(s) of variation can determine sampling intervals for ground truth gathering (field work to determine landform classification), placement of training points for geostatistically-derived supervised classification [4], filtered mapping of imagery[1], and model determination for various interpolation approaches. Fast computation of variograms is desirable for many military and environmental applications. If compressed images produce the same or nearly the same variogram as a full image, computational time can be reduced by several orders of magnitude. The experiment described here tests the quality of variograms of reduced images compared with the full images.

EXPERIMENTS AND ANALYSIS:

The data being analyzed in this paper is 1-meter, 4-band multispectral imagery of Fort A.P. Hill, Virginia. The imagery was collected by a Computerized Airborne Multicamera Imaging System (CAMIS) sensor flown from a Lear jet. The four bands are blue, green, red and near infrared. Imagery was taken in data frames of 768 x 576 pixels. Multiple flight lines (10 to 15) are taken for each collect, with 30 data frames in each flight line. Different frames were then registered to each other and mosaicked into one final image which is then radiometrically corrected and registered to existing maps.

Ft. A.P. Hill Mosaic

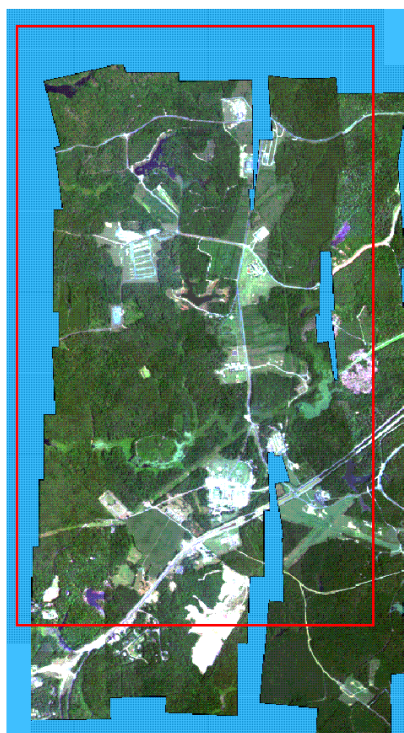


Figure 2: Imagery of Ft. A.P. Hill, VA used in the analysis

First, variograms of the full image were computed. Figure 3 shows the variogram for the full image of Band 1.

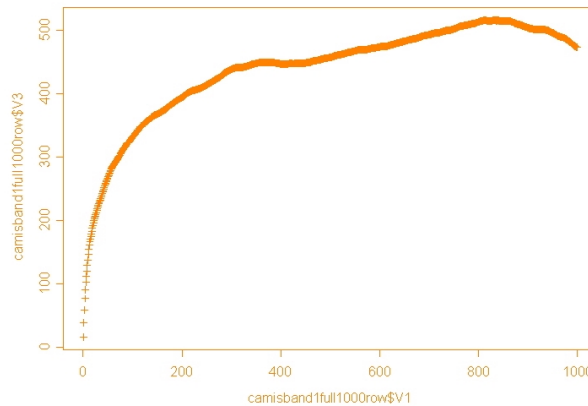


Figure 3: Variogram of Ft. A.P. Hill, VA CAMIS Image; full image, Band 1

The full image was then subsampled incrementally by $\frac{1}{4}$ down to a $\frac{1}{256}$ image, a reduction of two orders of magnitude. Variograms of the subsampled images were computed. No apparent difference between the variogram for the full image and the variograms for the $\frac{1}{256}$ image was visible. Figures 4 and 5 show the variogram of the $\frac{1}{256}$ image for Band 1, and the variograms for the full image and the $\frac{1}{256}$ image of Band 1 superimposed on each other.

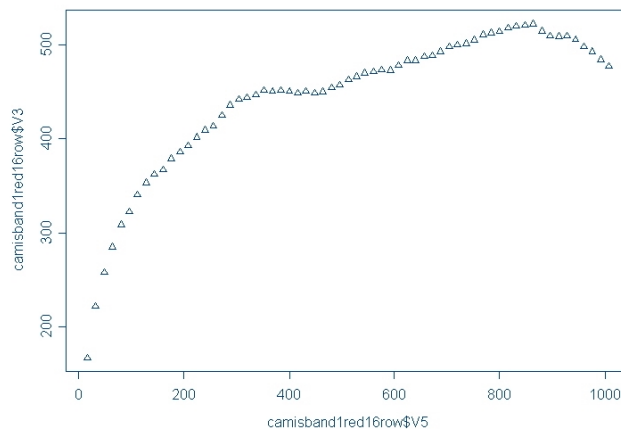


Figure 4: Variogram of Ft. A.P. Hill, VA CAMIS Image; 256X reduced image, Band 1

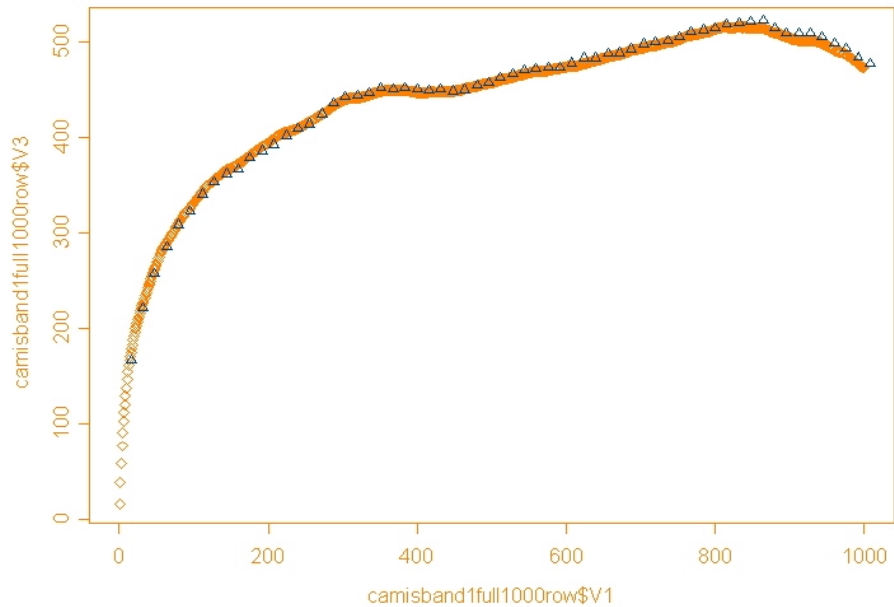


Figure 5: Variograms of Ft. A.P. Hill, VA CAMIS Image; full (orange) and 256X reduced (blue) images superimposed, Band 1

The two variograms are very close in shape; the resulting models are also very close, with spatial variation scales within a few meters of each other.

CONCLUSIONS AND FUTURE WORK:

Results show that imagery can be reduced by 256 without any appreciable degradation in the spatial dependence model. The variograms of the full image and of the reduced image are nearly identical, and the scales of variation resulting from modeling these two variograms are not different statistically. This seems to indicate that 1-meter imagery can be compressed by an order of two magnitudes (10^2) without affecting the results of spatial statistical analysis. This can reduce the time required to compute variograms of large 1-meter imagery sets from hours to minutes or seconds, which is substantially closer to real-time applications of spatial statistical imagery analysis.

Future work includes testing the effects of compression on other imagery to verify that this result holds in the general case.

REFERENCES:

- [1] Oliver, M.A., Webster, R., and Slocum, K., "Image Filtering by Kriging Analysis", International Journal of Remote Sensing, 1999.
- [2] Shine, "Mapping and Modeling 1-Meter Multispectral Imagery Data", proceedings of the Joint Statistical Meetings, Indianapolis, IN, August 2000.
- [3] Matheron, G., The Theory of Regionalized Variables and Its Applications, Fontainebleau, 1971.
- [4] Shine, J.A. and Wakefield, G.I., "A Comparison of supervised Imagery Classification Using Analyst-Chosen and Geostatistically-Chosen Training Sets", 4th International Conference on Geocomputation, Fredericksburg, VA, July 1999.