Geospatial Analytics for Big Spatiotemporal Data: Algorithms, Applications, and Challenges*

Ranga Raju Vatsavai† and Budhendra Bhaduri
Computational Sciences and Engineering Division
Oak Ridge National Laboratory, Oak Ridge, TN 37831.

ABSTRACT

Explosive growth in the spatial and spatiotemporal data and the emergence of social media and location sensing technologies emphasize the need for developing new and computationally efficient geospatial analytics tailored for analyzing big data. In this white paper, we review major spatial data mining algorithms by closely looking at the computational and I/O requirements and allude to few applications dealing with big spatial data.

1. DATA CHALLENGES

We are living in the era of 'Big Data.' Spatiotemporal data, whether captured through remote sensors (e.g., remote sensing imagery, Atmospheric Radiation Measurement (ARM) data) or large scale simulations (e.g., climate data) has always been 'Big.' However, recent advances in instrumentation and computation making the spatiotemporal data even bigger, putting several constraints on data analytics capabilities. In addition, large-scale (spatiotemporal) data generated by social media outlets is proving to be highly useful in disaster mapping and national security applications. Spatial computation needs to be transformed to meet the challenges posed by the big spatiotemporal data. Table 1 shows some of the climate and earth systems data stored at the Earth System Grid (ESG) portal.

	CMIP5	ARM	DACC
Sponsor	SciDAC	DOE/BER	NASA
Description of Data	40+ Models	Atmospheric Processes and Cloud Dynamics	Biogeo-chemical dynamics, FLUXNET
Archive Size	$\sim$ 6 PB	$\sim$ 200 TB	$\sim$ 1 TB
Year Started	2010	1991	1993

Table 1: ESG Integrated Data Archive‡

In addition to the data archived at ESG portal, remote sensing imagery data archived at the NASA EOSDIS exceeds 3 PB. NASA generates about 5 TB of data per day. Figure 1 shows progression of remote sensing instruments along three important sensor characteristics: spatial, spectral, and temporal resolutions. Though these improvements are leading to increase in volume, velocity, and variety of remote sensing data products and making it hard to manage and process, they are also enabling new applications. For example, improvements in temporal resolution allows monitoring biomass on a daily basis. Improvements in spatial resolution allows fine-grained classification (settlement types), damage assessments, and critical infrastructure (e.g., nuclear proliferation) monitoring.

Google generates about 25 PB of data per day, significant portion of which is spatiotemporal data (images and videos). The rate at which spatiotemporal data is being generated clearly exceeds our ability to organize and analyze them to extract patters critical for understanding dynamically changing world. Therefore, we need focused research on developing efficient management and analytical infrastructure for big spatial data. In this paper we review major spatial data mining algorithms and applications by closely looking at the computational and I/O challenges posed by the `big spatial data`.

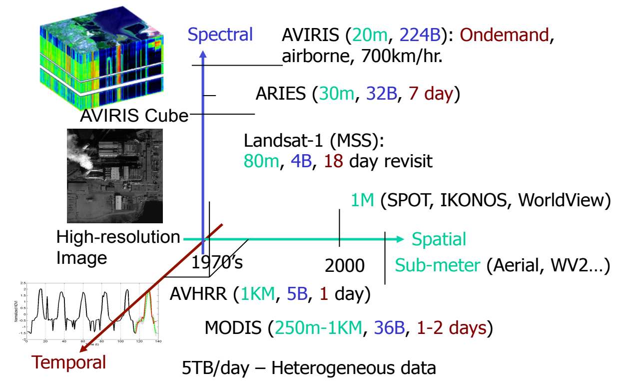

Figure 1: Advances in remote sensing data products (1970's through present)

2. COMPUTATIONAL CHALLENGES

Increasing spatial and temporal resolution requires that the data mining algorithms should take into account the spatial and temporal autocorrelation. Explicit modeling of spatial dependencies increase computational complexity. We

*Big Data and Extreme Scale Computing (BDEC'13), April 30 - May 1, Charleston, 2013. Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725. ‡(Table 1: Curtesy of G. Shipman)

†Corresponding author (vatsavairr@ornl.gov)

now briefly look at the following widely used data mining primitives that explicitly model spatial dependencies: spatial autoregressive (SAR) model, Markov Random Field (MRF) model, Gaussian Processes Learning and Mixture Models. More details about these techniques can be found in [1] and references therein.

1. **Spatial Autoregressive Model(SAR):** In prediction problems involving spatial data, often spatial dependencies are modeled in regression through a neighborhood matrix W as given by: $\mathbf{y} = \rho W \mathbf{y} + \mathbf{X}\beta + \epsilon$. The estimates of ρ and β can be derived using maximum likelihood theory or Bayesian statistics, which involves Markov Chain Monte Carlo (MCMC) sampling. Computational complexity of direct likelihood-based estimation is $O(n^3)$ and memory requirements is $= O(n^2)$.

2. **Markov Random Field Classifiers:** Spatial dependencies in classification are often modeled through the extension of *a priori* probabilities in a Bayesian classification framework as given by:

$$Pr(l(s_i)|X, L\backslash l(s_i)) = \frac{Pr(X(s_i)|l(s_i), L\backslash l(s_i)) Pr(l(s_i)|L\backslash l(s_i))}{Pr(X(s_i))}$$

   The solution procedure can estimate $Pr(l(s_i)|L\backslash l(s_i))$ from the training data by examining the ratios of the frequencies of class labels to the total number of locations in the spatial framework. $Pr(X(s_i)|l(s_i), L\backslash l(s_i))$ can be estimated using kernel functions from the observed values in the training dataset. The solution procedure involves costly iterative optimizations or graph cuts.

3. **Gaussian Process (GP) Learning:** Modeling spatial heterogeneity is also important in classification of large geographic regions. GP learning extends Bayesian classification framework through class-conditional distribution, where any given i-th class is modeled as a function of spatial coordinate $\mathbf{s}$: $p(\mathbf{x}(\mathbf{s})|y_i) \sim N(\boldsymbol{\mu_i}(\mathbf{s}),\ \Sigma_i)$. GP based learning can also be used for change detection. Recently we extended this model for biomass monitoring through an exponential periodic covariance function, which leads to special Toeplitz matrix. Where as the computational complexity of original solution of GP learning is $O(n^3)$, Toeplitiz matrix representation requires only $O(n)$ memory and inversion requires only $O(n^2)$ complexity. Even after employing computationally efficient algorithms, change detection is a challenging task. For example, in the case of biomass monitoring using coarse spatial resolution (250 meters) MODIS data, one has to process 23,040,000 time series for one (tile) image. One has to process 326 such tiles (that is, 7,511,040,000 individual time series) in a day before new images arrive.

3. APPLICATIONS

In this section we present a diverse but representative set of applications which are dealing with big spatial data.

1. **Biomass Monitoring:** Monitoring biomass over large geographic regions for identifying changes is an important task in many applications. Monitoring biomass over a large geographic region requires high temporal resolution satellite imagery. NASA's Terra satellite with the MODIS instrument aboard is providing a new opportunity for continuous monitoring of biomass over large geographic regions. Since data at global scale is difficult to handle, MODIS data is organized into tiles of $10^o \times 10^o$ (4800 x 4800 pixels). Though there are 460 daily MODIS tile products available, we need to process 326 products which contain land pixels. At daily temporal resolution, MODIS time series contains about 3600 data points (at each pixel location). Often the computational complexity of change detection algorithms is very high, for example, GP learning presented in the Section **??** is $O(n^3)$ and $O(n^2)$, where n is the number of data points in each time series. In addition we need to process about 7,511,040,000 time series in a day, where each time series contains 3600 data points, before new set of MODIS data products arrive.

2. **Searching for Complex Patterns:** Most of the pattern recognition and machine learning algorithms are per-pixel based (or single instance). These methods worked well for thematic classification of moderate and high-resolution (5 meters and above) images. Very high resolution (VHR) images (sub-meter) are offering new opportunities beyond thematic mapping, they allow recognition of structures in the images. Example applications include: (1) recognizing complex spatial patterns in an urban setting to map informal (slums) and formal settlements, (2) recognizing critical infrastructure (e.g., nuclear, thermal, and chemical plants, airports, shopping and sports complexes), and (3) image based content search and retrieval. These tasks require feature extraction and selection, indexing, machine learning, and pattern matching. Many of these tasks often deal with segments or objects as opposed to pixels. Computing match (similarity) between image patches (e.g., Housdorff distance) is computationally expensive, often $O(n^2)$. Therefore, scaling these algorithms for global applications requires not only efficient and novel algorithmic solutions, but also require exascale computing infrastructure to support global spatiotemporal applications.

4. CONCLUSIONS

Big spatiotemporal data is supporting key applications of national importance like food, energy, and national security. However, in order to realize the full benefits of big spatiotemporal data, one has to overcome both computational and I/O challenges. We not only need new models that explicitly model spatial and temporal constraints efficiently, but further research is also required in the area of compression, sampling, approximate solutions that guaranty quality but are orders of magnitude faster, algorithms that deal with heterogeneity (multiple sensors, multiple resolutions), numerical optimization, and algorithms for heterogenous architectures.

5. REFERENCES

[1] R. R. Vatsavai and et. al. Spatiotemporal data mining in the era of big spatial data: algorithms and applications. In *ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, BigSpatial '12. ACM, 2012.