Spatiotemporal Data Mining in the Era of Big Spatial Data: Algorithms and Applications*

Ranga Raju Vatsavai†
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831.
vatsavairr@ornl.gov

Varun Chandola
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831.
chandolav@ornl.gov

Scott Klasky
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN 37831.
klasky@ornl.gov

Auroop Ganguly
Department of Civil and Environmental Engineering
Northeastern University
Boston, Massachusetts 02115
a.ganguly@neu.edu

Anthony Stefanidis
Department of Geography and Geoinformation Science
George Mason University
MS 6C3, Fairfax VA 22030
astefani@gmu.edu

Shashi Shekhar
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455
shekhar@cs.umn.edu

ABSTRACT

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from the spatial and spatiotemporal data. However, 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 methods tailored for analyzing big data. In this 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.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Learning; I.5 [**Pattern Recognition**]: Classification, Clustering, Prediction

General Terms

Big Data, Large Scale Data Mining, Spatiotemporal Patterns, Computational and I/O Challenges

†Corresponding author.

ACM BigSpatial '12 November 6-9. 2012. Redondo Beach, CA, USA

1. INTRODUCTION

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. 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	Biogeochemical 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`.

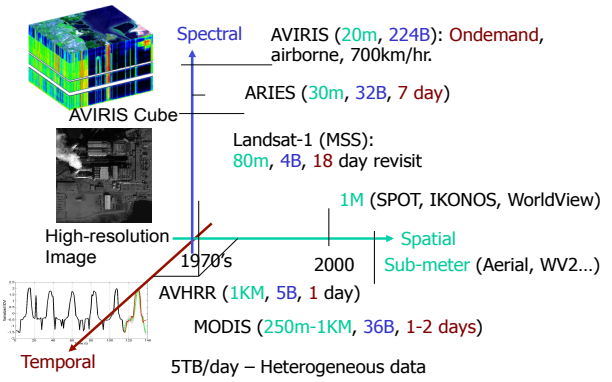


Figure 1: Advances in remote sensing data products (1970's through present)

2. ALGORITHMS

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 now briefly look at the following widely used data mining primitives that explicitly model spatial dependencies: spatial autoregressive (SAR) model [2, 38, 37, 47, 48, 39, 42], Markov Random Field (MRF) model [23, 14, 27, 54, 6, 61, 40, 1], Gaussian Processes [49], and Mixture Models [44]. More details about these techniques can be found in [58] and references therein.

2.1 Spatial Autoregressive Model(SAR)

We now show how spatial dependencies are modeled in the framework of regression analysis. In spatial regression, the spatial dependencies of the error term, or, the dependent variable, are directly modeled in the regression equation[2]. If the dependent values y_i' are related to each other, i.e., $y_i = f(y_j)$ $i \neq j$, then the regression equation can be modified as

$$\mathbf{y} = \rho W\mathbf{y} + \mathbf{X}\beta + \epsilon. \quad (1)$$

Here W is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of spatial dependencies between the elements of the dependent variable. After the correction term $\rho W\mathbf{y}$ is introduced, the components of the residual error vector ϵ are then assumed to be generated from independent and identical standard normal distributions.

Computational and I/O Challenges

The estimates of ρ and β can be derived using maximum likelihood theory or Bayesian statistics. Bayesian approach using sampling-based Markov Chain Monte Carlo (MCMC) methods can be found in [37]. Without any optimization, likelihood-based estimation would require $O(n^3)$ operations. Recently [47, 48, 39, 9] have proposed several efficient techniques to solve SAR. Many of these techniques have been studied and compared in [34]. SAR model also requires an $= O(n^2)$ memory to store the neighborhood matrix $\mathbf{W}$.

2.2 Markov Random Field Classifiers

A set of random variables whose interdependency relationship is represented by a undirected graph (i.e., a symmetric neighborhood matrix) is called a Markov Random Field [40]. The Markov property specifies that a variable depends only on the neighbors and is independent of all other variables. The location prediction problem can be modeled in this framework by assuming that the class label, $f_L(s_i)$, of different locations, s_i, constitute an MRF. In other words, random variable $f_L(s_i)$ is independent of $f_L(s_j)$ if $W(s_i, s_j) = 0$. The Bayesian rule can be used to predict $f_L(s_i)$ from feature value vector X and neighborhood class label vector L_M as follows:

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

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. For reliable estimates, even larger training datasets are needed relative to those needed for the Bayesian classifiers without spatial context, since we are estimating a more complex distribution. This is accomplished by taking contiguous regions (windows) instead of sample points. An assumption on $Pr(X(s_i)|l(s_i), L\backslash l(s_i))$ may be useful if large enough training data set is not available. A common assumption is the uniformity of influence from all neighbors of a location. Another common assumption is the independence between X and L_N, hypothesizing that all interaction between neighbors is captured via the interaction in the class label variable. Many domains also use specific parametric probability distribution forms, leading to simpler solution procedures. In addition, it is frequently easier to work with the Gibbs distribution specialized by the locally defined MRF through the Hammersley-Clifford theorem [3].

Computational and I/O Challenges

Solution procedures for the MRF Bayesian classifier include stochastic relaxation [23], iterated conditional modes [4], dynamic programming [18], highest confidence first [14] and Graph cut [6]. We have explored the graph cut method in the past, more details can be found in [53].

2.3 Gaussian Process Learning

MRFs described in the previous section are widely used to model spatial homogeneity. However, modeling spatial

heterogeneity is also important in classification. Statistical modeling of spatial variation has been well known as spatial statistics or geostatistics. The process of finding the optimal linear predictor is called kriging, named after a South African mining engineer, D. G. Krige [15]. In the machine learning community, the same model is known as the Gaussian process regression model. When the underlying stochastic process is a Gaussian random process, the linear predictor obtained by kriging is optimal in least-square sense. More specifically, the Gaussian process regression model corresponds to a simple or ordinary kriging model.

The conventional maximum-likelihood classifier (MLC) typically models the class-conditional distribution, $p(\mathbf{x}|y)$, as a multi-variate Gaussian distribution:

$$p(\mathbf{x}|y = y_i) \sim N(\boldsymbol{\mu_i}, \Sigma_i) \ , \tag{2}$$

where $\mathbf{x} = (x_1, x_2, ..., x_d)^T$ is a d-dimensional vector representing spectral bands of a pixel in a hyperspectral image, and $y \in \{y_1, y_2, ..., y_c\}$ is the LULC class label. The parameters for multi-variate Gaussians, $\boldsymbol{\theta_i} = (\boldsymbol{\mu_i}, \Sigma_i)$, are obtained by the maximum-likelihood estimation (MLE), and assumed to be constant over all possible locations. As discussed earlier, this assumption do not hold in general. Spatial heterogeneity can be modeled via non-parametric Gaussian process model. In this model, the class-conditional distribution of the i-th class is modeled as a function of spatial coordinate $\mathbf{s}$:

$$p(\mathbf{x}(\mathbf{s})|y_i) \sim N(\boldsymbol{\mu_i}(\mathbf{s}),\ \Sigma_i) \ . \tag{3}$$

where $\boldsymbol{\mu_i(s^*)} = (\mu_{i1}(\mathbf{s}^*), \mu_{i2}(\mathbf{s}^*), ..., \mu_{id}(\mathbf{s}^*))$. We omit i that indicates the i-th class in the following equations to simplify the notation. Each spectral band of $\mathbf{x}$ is modeled as a random process indexed by a spatial coordinate $\mathbf{s} = (s_1, s_2)$, then the j-th band of $\mathbf{x}$, x_j, can be written as

$$x_j(\mathbf{s}) = f_j(\mathbf{s}) + \epsilon_j \ , \tag{4}$$

where $f_j(\mathbf{s})$ is a Gaussian random process and ϵ_j is an additive white Gaussian noise (AWGN):

$$\epsilon \sim \mathcal{N}(0, \sigma^2_{\epsilon_j}) \ .$$

Given $f_j(\mathbf{s})$, then the class conditional distribution of x_j is

$$p(x_j(\mathbf{s})|f_j(\mathbf{s})) = \mathcal{N}(f_j(\mathbf{s}),\ \sigma^2_{\epsilon_j}) \ .$$

We assume a (zero-mean) Gaussian process for $f_j(\mathbf{s})$:

$$f_j(\mathbf{s}) \sim \mathcal{GP}(0,\ K_j(\mathbf{s}_l, \mathbf{s}_m)) \ ,$$

where $K_j(\mathbf{s}_l, \mathbf{s}_m)$ is a spatial covariance function between locations $\mathbf{s}_l$ and $\mathbf{s}_m$. The zero-mean prior assumption correspond to the simple kriging model in spatial statistics [15]. In practice, we can approximately satisfy the zero-mean assumption by normalizing given feature values. Characteristics of a Gaussian random process is solely defined by a covariance function. More details on GP-based classification can be found in [28]. In addition to the classification, we have adopted GP for biomass change detection over large areas [10]. For change detection, we used a covariance function known as *Exponential Periodic (ep)*:

$$k(t_1, t_2) = \sigma_f^2 exp(-\frac{\Delta t^2}{2l^2\omega^2}) exp(-\frac{(1 - \cos\frac{2\pi\Delta t}{\omega})}{a}) \tag{5}$$

where ω is the length of a single cycle of the periodic time series. This covariance function effectively models the periodic time series.

$$K = \begin{pmatrix} k_0 & k_1 & k_2 & \dots & k_{t-1} \\ k_1 & k_0 & k_1 & \dots & k_{t-2} \\ k_2 & k_1 & k_0 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ k_{t-1} & k_{t-2} & \dots & \dots & k_0 \end{pmatrix} \tag{6}$$

Computational and I/O Challenges

It is well-known that GP do not scale well for large datasets. First, the covariance function (e.q. 6) requires $O(n^2)$ memory. The solution consists of estimating the following terms: $\mathbf{y}^\mathsf{T}K^{-1}\mathbf{y}$, $k^\mathsf{T}K^{-1}\mathbf{y}$, $k^\mathsf{T}K^{-1}k$. Straight forward solutions are computationally expensive, $O(t^3)$. However, noting that the covariance function 6 leads to a symmetric Toeplitz and positive semi-definitive matrix, the memory and computational requirements reduces to $O(n)$ and $O(n^2)$ respectively. More details can be found in [10]. 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.

2.4 Mixture Models

Mixture models are widely used in clustering and semi-supervised learning. In this section we present a Gaussian Mixture Model (GMM) based clustering algorithm. GMM based clustering consists of two subproblems. First, we have to estimate the model parameters. Second, we need to estimate the number of components in the GMM.

2.4.1 Estimating the GMM Parameters

First we solve the model parameter estimation problem by assuming that the training dataset D_j is generated by a finite Gaussian mixture model consisting of M components. If the labels for each of these components were known, then problem simply reduces to usual parameter estimation problem and we could have used MLE. We now describe a parameter estimation technique that is based on the well-known expectation maximization algorithm. Let us assume that each sample x_j comes from a super-population D, which is a mixture of a finite number (M) of clusters, $D_1, \dots, D_M$, in some proportions $\alpha_1, \dots, \alpha_M$, respectively, where $\sum_{i=1}^{M} \alpha_i = 1$ and $\alpha_i \geq 0 (i = 1, \dots, M)$. Now we can model the data $D = \{x_i\}_{i=1}^n$ as being generated independently from the following mixture density.

$$p(x_i|\Theta) = \sum_{j=1}^{M} \alpha_j p_j(x_i|\theta_j) \tag{7}$$

$$L(\Theta) = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{M} \alpha_j p_j(x_i|\theta_j) \right] . \tag{8}$$

Here $p_j(x_i|\theta_j)$ is the pdf corresponding to the mixture j and parameterized by θ_j, and $\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$ denotes all unknown parameters associated with the M-component mixture density. The *log-likelihood* function for

this mixture density is given in 8. In general, Equation 8 is difficult to optimize because it contains the ln of a sum term. However, this equation greatly simplifies in the presence of unobserved (or incomplete) samples. We now simply proceed to the expectation maximization algorithm, interested reader can find detailed derivation of parameters for GMM in [5]. The expectation maximization (EM) algorithm at the first step maximizes the expectation of the *log-likelihood* function, using the current estimate of the parameters and conditioned upon the observed samples. In the second step of the EM algorithm, called maximization, the new estimates of the parameters are computed. The EM algorithm iterates over these two steps until the convergence is reached. For multivariate normal distribution, the expectation $E[.]$, which is denoted by p_{ij}, is the probability that Gaussian mixture j generated the data point i, and is given by:

$$p_{ij} = \frac{\left|\hat{\Sigma}_j\right|^{-1/2} e^{\left\{-\frac{1}{2}(x_i-\hat{\mu}_j)^t \hat{\Sigma}_j^{-1}(x_i-\hat{\mu}_j)\right\}}}{\sum_{l=1}^{M}\left|\hat{\Sigma}_l\right|^{-1/2} e^{\left\{-\frac{1}{2}(x_i-\hat{\mu}_l)^t \hat{\Sigma}_l^{-1}(x_i-\hat{\mu}_l)\right\}}} \quad (9)$$

The new estimates (at the k^{th} iteration) of parameters in terms of the old parameters at the M-step are given by the following equations:

$$\hat{\alpha}_j^k = \frac{1}{n}\sum_{i=1}^{n} p_{ij} \quad (10) \qquad \hat{\mu}_j^k = \frac{\sum_{i=1}^{n} x_i p_{ij}}{\sum_{i=1}^{n} p_{ij}} \quad (11)$$

$$\hat{\Sigma}_j^k = \frac{\sum_{i=1}^{n} p_{ij}(x_i-\hat{\mu}_j^k)(x_i-\hat{\mu}_j^k)^t}{\sum_{i=1}^{n} p_{ij}} \quad (12)$$

2.4.2 Clustering

Once the GMM is fitted to the training data, we can use the model to predict labels for each cluster. The assignment of label is carried out using the maximum likelihood (ML) procedure. The discriminant function g(.) given by ML principle is as following:

$$g_i(x) = -\ln|\Sigma_i| - (x-\mu_i)^t|\Sigma_i|^{-1}(x-\mu_i) \quad (13)$$

For each pixel (feature vector), we assign a cluster label i, if $g_i(x)$ is maximum over all cluster labels.

Computational and I/O Challenges

The computational complexity of GMM model fitting depends on the number of iterations and time to compute expectation (E) and maximization (M) steps. Let us assume that size of training dataset size is N, and the number of components is M, and the dimensionality is d. Then the cost of E- and M-steps are $O(NMD+NM)$ and $O(2NMD)$ respectively at each iteration. On the other hand spatial extension [60] incurs additional cost of iterative conditional mode (ICM) step. We developed an efficient solution for GMM based spatial semi-supervised learning in [60]. We have also parallelized GMM clustering algorithm on GPUs. Initial results on GTX285 with 240 CUDA cores and 1GB memory shows excellent scalability of 160x on learning part. Learning part is typically computationally expensive and less I/O intensive, as we we have to deal with small training data which is typically 3-5% of the total data. However, for clustering (e.q. 13), that is assigning label to each pixel in the image, the performance is suffered by I/O as we have to deal with 95-97% of the total data. The main reason being that the computational requirements of clustering are modest (e.q. 13), as compared to the learning, but the number of samples to process are huge. We need efficient I/O schemes to scale-up clustering for large datasets.

3. APPLICATIONS

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

3.1 Biomass Monitoring

Monitoring biomass over large geographic regions for identifying changes is an important task in many applications. With recent emphasis on biofuel development for reducing dependency on fossil fuels and reducing carbon emissions from energy production and consumption, the landscape of many countries is going to change dramatically in coming years. Already there are several preliminary reports that address both economic and environmental impacts of growing energy crops. In the United States continuous corn production is becoming a dominant cropping pattern as more and more soybean and wheat rotations are replaced by continuous corn production. It is also expected that more and more pasture lands will be converted to Switchgrass in the coming years, which may positively impact climate change because of its superior carbon uptake properties. These changes are not limited to the United States alone. Developing countries like India, the rural areas are facing increasing demand for energy. It is expected that energy crops like Jatropha curcas are going to be widely planted in Asian countries. Recent FAO report [7] indicates a threefold increase in the area planted to jatropha from 4.72 million ha in 2010 to 12.8 million ha by 2015.

Monitoring biomass over a large geographic region requires high temporal resolution satellite imagery. The launch of NASA's Terra satellite in December of 1999, with the MODIS instrument aboard, introduced a new opportunity for continuous monitoring of biomass over large geographic regions. MODIS data sets represent a new and improved capability for terrestrial satellite remote sensing aimed at meeting the needs of global change research [29]. MODIS land products are generally available within weeks or even days of acquisition and distributed through the EROS Data Center (EDC) and are currently available free of charge. The availability of multi-temporal MODIS imagery has made it possible to study plant phenology, quantitatively describe NPP patterns in time and space, and monitor and map natural resources at regional and global scales. MODIS allows users to identify vegetation changes over time across a region and estimate quantitative biophysical parameters which can be incorporated into global climate models. Even though several cumulative vegetation indices can be found in the literature, MODIS NDVI temporal profiles are widely used in studying plant phenology.

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 2.3 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. In addition to finding changes, we also need to characterize changes using high-resolution satellite image products which put tremendous constraints on computational resources. More details about a scalable biomass monitoring system can be found in [57, 12, 11].

3.2 Complex Object Recognition

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. For example, consider the problem of settlement mapping [25]. The high rate of urbanization, political conflicts and ensuing internal displacement of population, and increased poverty in the 20th century has resulted in rapid increase of informal settlements. These unplanned, unauthorized, and/or unstructured homes, known as informal settlements, shantytowns, barrios, or slums, pose several challenges to the nations as these settlements are often located in most hazardous regions and lack basic services. Though several World Bank and United Nations sponsored studies stress the importance of poverty maps in designing better policies and interventions, mapping slums of the world is a daunting and challenging task. VHR images provides the ability to distinguish informal settlements from formal settlements. However, per-pixel based methods do not work well for very high-resolution (VHR) images (submeter). The main problem being that the pixel size (less than meter) is too small as compared to the object size (10s of meters) and contains too little contextual information to accurately distinguish between given set of pixels. As shown in Figure 2 often do not provide sufficient discrimination power between classes. One way to alleviate this problem is to consider a bigger window or patch consisting a group of adjacent pixels which offers better spatial context than a single pixel. Unfortunately, this makes all well known per-pixel based classification schemes ineffective. Multi-instance learning approaches might be useful in moving from pixel-based or object-based structure recognition in VHR images, but computational complexity is too high to be practically applied for global settlement mapping.

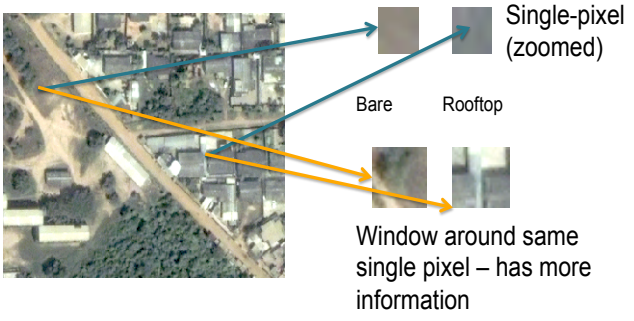


Figure 2: Problems with pixel-based pattern recognition methods

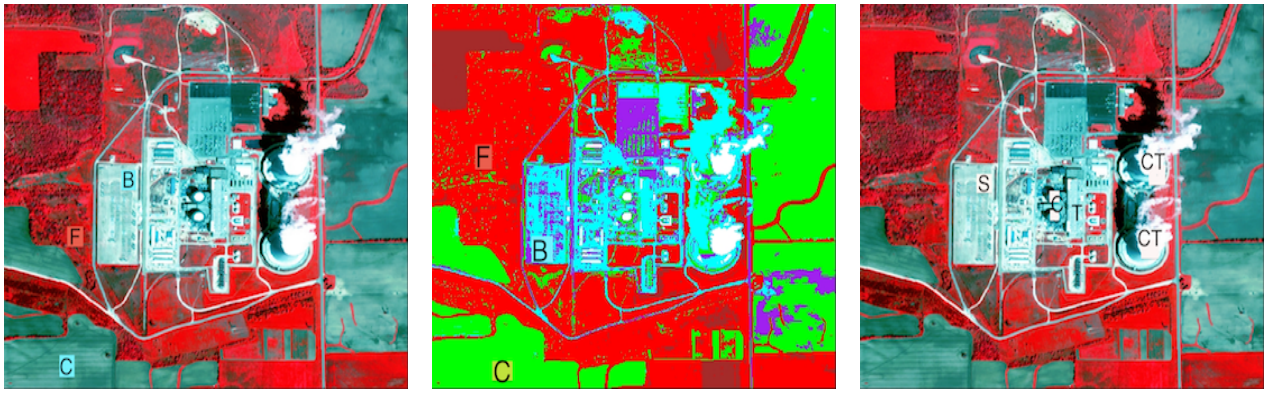
Now, let us consider the problem of identifying complex facilities (e.g., nuclear facilities, thermal power plants) [59] in VHR images. As can be seen from Figure 3, thematic classification is designed to learn and predict thematic classes such as forest (F), crops (C), buildings (B), etc., at pixel level. However, such thematic labels are not enough to capture the fact that the given image contains a nuclear power plant. What is missing is the fact that the objects, such as switch yard (S), containment building (C), turbine building (T), and cooling towers (CT) have distinguishing shapes, sizes, and spatial relationships (arrangements or configurations) as shown in Figure 3(c). These semantics are not captured in the traditional pixel- and object-based classification schemes. In addition, traditional image analysis approaches mainly exploit low-level image features (such as, color and texture and, to some extent, size and shape) and are oblivious to higher level descriptors and important spatial (topological) relationships without which we can not accurately discover these complex objects or higher level semantic concepts. Figure 4 shows four different images (baseball and football fields, two residential neighborhoods) where they share common objects, for example, grass and soil across baseball and football fields, and two (economically) different neighborhoods where in one neighborhood buildings are colocated with cars (parked on the road) while in the other builds are collocated with swimming pools. Both pixel- and object-based methods often fails to capture these complex relationships. Future research requires models that explicitly learn complex spatial relationships among the objects to accurately predict semantic classes and scale to big VHR image collections.

3.3 Climate Change Studies

The big data challenge is rising to be among the major concerns in furthering our understanding of climate science and translating the science to information relevant for impacts and policy. Here we show, through specific examples drawn primarily from our prior work, how the complexity, size, and heterogeneity of the climate data and corresponding solutions from the data and computational sciences have the potential to transform the field.

The Intergovernmental Panel on Climate Change (IPCC), in their Nobel Prize winning Fourth Assessment Report (AR4) in 2007, primarily used simulations from the Coupled Model Intercomparison Project Phase 3 (CMIP3), in addition to observations, to support two major hypotheses: First, global warming is real, and second, human emissions of greenhouse gases are to blame. The CMIP3 multi-model archives, which started at tens of terabytes, have grown to a few hundred terabytes since. The latest generation, or CMIP5, model archives have already reached the petabyte scale and are expected to grow to beyond a couple of petabytes. The size of remotely sensed data is beginning to catch up as well. According to a recent article [46] in Science magazine, climate-related data is expected to reach about 350 petabytes by the 2030s, with about half of that contributed by models, a bit less than half by remote sensing data, and the rest from in situ sensors. The focus of climate science, meanwhile, has been shifting more towards the regional or even local scales, with a view to consequence assessments and informing adaptation decisions. This has in turn motivated regional climate modeling runs as well as statistical downscaling, which could result orders of magnitude increase in data size.

In addition to the sheer size, the complexity of climate



(a) FCC Image with Thematic class labels (B-Buildings, C-Crop, F-Forest)

(b) Thematic Classified Image (B-Buildings, C-Crop, F-Forest)

(c) FCC Image with Semantic Labels (S-Switch Yard, C-Containment Building, T-Turbine Generator, CT-Cooling Tower

Figure 3: Thematic vs. Semantic Classes

data contributes to computational and data science challenges [21]. While data mining or even machine learning methods and principles were originally developed for independent and identically distributed data, the areas of time series, spatial and spatiotemporal data mining partly emerged as a response to the growing challenge of finding interesting and novel patterns from correlated but massive data. Time series analysis and forecasting in statistics typically starts with an examination of the dependence structure, for example, through the autocorrelation function, while signal processing starts with the power spectrum which is a representation of the dependence structure in frequency space and hence a Fourier transform of the autocorrelation function. Spatial and spatiotemporal statistics rely on the fundamental notion that values closer to each other in time or space are more closely related to each other, with due considerations for seasonality and low frequency components. In fact, the first law of geography states that all things being equal things closer to each other are more closely related. The fundamental principles of time series, spatial and spatiotemporal data mining often rely on the same basic principles, for example, auto- or cross-correlation, in space and/or time. Thus, regression and weighted regression techniques need to be adapted through geographical weights or spatial and temporal constraints, leading to numerical and computational challenges. For climate data and processes however, the situation grows worse. First, the dynamical climate system exhibits interconnections and dependence structures that may often only be observed as teleconnections and captured through nonlinear correlation measures [35] or graphical dependence structures [33], including complex networks [56] or probabilistic graphs [13]. Second, climate processes operate at multiple scales, from the intraseasonal scales of the Madden-Julian Oscillation (MJO) to the interannual El Nino Southern Oscillation (ENSO) all the way to the Atlantic Multidecadal Oscillation (AMO), just as a few examples. The processes can vary across not just time but also spatial scales, which enhance the challenge. From a computational and data perspective, the multiscale behavior of the climate system makes the combinatorics for dependence and prediction structures significantly more complex. Third, climate processes are nonlinear dynamical, and may even exhibit chaos or extreme sensitivity to initial conditions, while the variability in the data may be non-Gaussian, even showing Ò1/fÓ noise. Fourth, climate projections and attributions are typically desired for the longer-term (decadal to century scales), which implies uncertainties owing to socioeconomic changes, technological policies, and mitigation regulations. Fifth, the problems of interest in climate are typically changes at regional scales or for the extremes [17, 30], which makes the data and computational challenges harder. Sixth, uncertainty quantification becomes a major challenge [22] owing to non-stationarity, nonlinear dynamics and long projection lead times with high precision. Finally, the consequence of climate change are felt on critical infrastructures and key resources, which implies the need to bring together climate analysis results together with impact-relevant data, for example, through geographical information systems. As climate related hazards and impacts on natural and human systems become more common, the detection of change under non-stationarity and nonlinearity becomes a major concern. Thus, civil and water resources engineers, who work with intensity-duration-frequency (IDF) curves for designing hydraulic infrastructures or for water resources decisions, would need to consider the changes on the very basis of their design or planning decisions [30].

Computational data sciences can help the science of climate change and impacts in multiple ways. Climate science can benefit from pattern discovery process assessment, analysis and uncertainty characterization, as well as enhanced predictions, particularly at regional scales and for extremes. The translation to impacts, adaptation and vulnerability can especially benefit from precise projections with comprehensive uncertainty characterizations.

3.4 Social Media Mining

In addition to the big data challenges associated with the proliferation of traditional geospatial datasets like remote sensing imagery, a new challenge is presented to the geoinformatics community through the emergence of novel types of geospatial datasets. For example, the massive adoption of large-scale video surveillance is imposing some novel challenges due to the exponential growth of the resulting

datasets. As a telling reference, in the UK alone, it is estimated that between 2 and 4 million CCTVs are deployed, with over 500,000 of them operating in London [45]. However, while the computational challenges associated with processing, storing, and analyzing such massive video datasets are substantial, one could still argue that they represent a slight evolution of traditional geospatial datasets. The same argument though cannot be made for social media feeds in the context of geospatial analysis, as they represent a substantial deviation from the traditional datasets that our community has been handling.

The geographic content of social media feeds represents a new type of geographic information. It does not fall under the established geospatial community definitions of crowdsourcing [20] or volunteered geographic information [24] as it is not the product of a process through which citizens explicitly and purposefully contribute geographic information to update or expand geographic databases. Instead, the type of geographic information that can be harvested from social media feeds can be referred to as Ambient Geographic Information (AGI) [55] it is embedded in the content of these feeds, often across the content of numerous entries rather than within a single one, and has to be somehow extracted. Nevertheless, it is of great importance as it communicates in real-time information about emerging issues, and also provides an unparalleled view of the complex social networking and cultural dynamics within a society, and captures the temporal evolution of the human landscape. Recent studies have shown how social media content can be harvested to monitor accurately the impact of natural disasters [16] or the spatiotemporal dynamics of a civil protest [62] The primary challenges imposed by such feeds relate to their volumes and heterogeneity.

Volume Considerations: Today, social media applications thrive. In the spring of 2011, just five years after its 2006 launch, twitter announced that it had over 200 million accounts, distributed all over the world. Among these accounts, it is estimated that twitter has at least 100 million active users, logging in at least once a month, and 50 million users who do so daily. As a measure of reference, a population of 100 million would make twitter the 12th most populous country in the world, just behind Mexico. Furthermore, twitterÕs *#numbers* entry indicates that a record 572,000 new accounts were created on a single day (March 12, 2011, the day after the Sendai earthquake and resulting nuclear disaster), while an average of 140 million tweets are sent daily, resulting in a billion tweets sent every week. However, it is not just twitter that has a large user community, facebook has reached 800 million users by the end of 2011, with more than half of them being active participants, using it daily. A population of 800 million would make facebook the third most populous nation in the world, behind only China and India. Extending beyond the English speaking word, QQ is a Chinese service for instant messaging, with over 800 million accounts, while Renren and Kaixin001 are the dominant Chinese social networking applications.

In addition to constantly increasing user communities, the amount of data released through social media applications is also increasing at very impressive rates. Seven years after its 2004 launch, flickr reached this year a landmark, by hosting 6 billion photos uploaded by its user community , with over 3,000 photos uploaded to flickr every minute and a 20% annual rate increase over the past few years. And while 6 billion photos is a very impressive number, it only reflects the estimated number of photos uploaded monthly to facebook by its user community, bringing the total number of photos hosted by facebook to nearly 100 billion . Regarding video, YouTube receives 48 hours of video uploads every minute, or the equivalent of 8 years of constant viewing content uploaded daily .

Data Heterogeneity: As social media comprise numerous platforms their content tends to be exceedingly diverse, ranging from text to photos and imagery. As a result, the form of raw geosocial media tends to be unstructured or ill-defined, and valuable knowledge is often hidden and cannot be easily processed through automation [51]. For example, both Twitter and Flickr provide APIs to query their content, but their responses are often structurally incompatible, not only between services but also within a single service, with incompatible schemas. Twitter can return the same content in different formats depending on the particular API used. Such data structure heterogeneity has a direct impact on the ability to store it in a single integrated database, and manage or process effectively. This heterogeneity is accompanied by a thematic diversity: user activity in Twitter can range for example from daily chatter and conversations, to reporting news, sharing information, or seeking information [26]. Furthermore, the narrative of these feeds is not simply a communication of information, but also reflects the contributorÕs perception of conventions established within the corresponding community: hashtag usage is Twitter is a perfect example of this situation [50]. Accordingly, the ability to integrate content across different social media feeds presents a challenge that requires a holistic collaborative approach to the analysis and synthesis of such data.

3.5 Mobility Applications

Mobility services, e.g., routing and navigation, are a set of ideas and technologies that transform lives by understanding the geo-physical world, knowing and communicating relations to places in that world, and navigating through those places. Mobility in this context can be defined as efficient, safe and affordable travel in our cities and towns. The transformational potential of mobility services is already evident. From Google Maps to consumer Global Positioning System (GPS) devices, society has benefited immensely from routing services and technology. Scientists use GPS to track endangered species to better understand behavior, and farmers use GPS for precision agriculture to increase crop yields while reducing costs. We have reached the point where a hiker in Yellowstone, a biker in Minneapolis, and a taxi driver in Manhattan know precisely where they are, their nearby points of interest, and how to reach their destinations.

We believe that harnessing big spatial data represents the next generation of routing services. Relevant examples include: temporally detailed (TD) roadmaps that provide speeds every minute for every road-segment, GPS trace data from cell-phones, and engine measurements of fuel consumption, greenhouse gas (GHG) emissions, etc. A 2011 McKinsey Global Institute report estimates savings of “about $600 billion annually by 2020” in terms of fuel and time saved [43] by helping vehicles avoid congestion and reduce idling at red lights or left turns. Preliminary evidence for the transformative potential includes the experience of UPS, which saves millions of gallons of fuel by simply avoiding left turns and associated engine-idling when selecting routes [41]. Im-

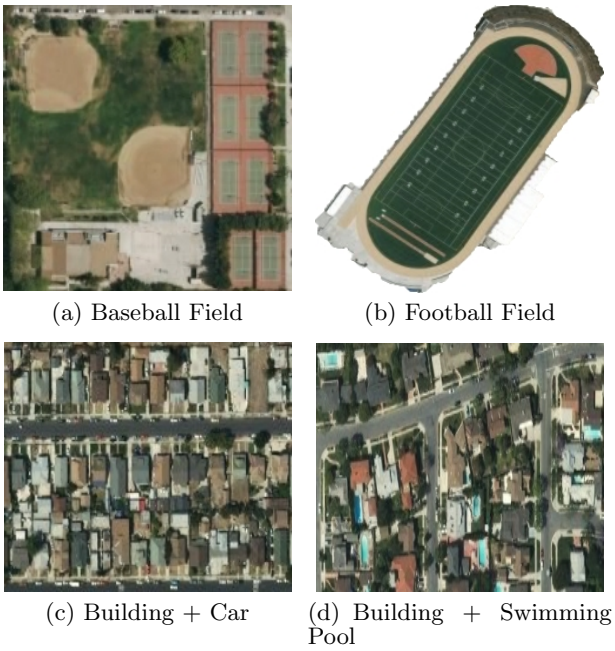
mense savings in fuel-cost and GHG emission are possible in the future if other fleet owners and consumers avoided left-turns and other hot spots of idling, low fuel-efficiency, and congestion. Eco-routing has the potential to significantly reduce US consumption of petroleum, the dominant source of energy for transportation. It may even reduce the gap between domestic petroleum consumption and production, helping bring the nation closer to the goal of energy independence.

There are two routing services that are widely in use. The first deals with determination of a best route given a start location, end location, optional waypoints, and a preference function. Here, choice of preference function could be: fastest, shortest, easiest, pedestrian, public transportation, avoid locations/areas, avoid highways, avoid tollways, avoid U-turns, and avoid ferries. There are several route finding solutions which are mostly based on classic shortest path algorithms, such as Dijktra and A*. Shortest path finding is often of interest to tourists as well as drivers in unfamiliar areas. In contrast, commuters often know a set of alternative routes between their home and work. They often use an alternate service to compare their favorite routes using real-time traffic information, e.g., scheduled maintenance and current congestion. These widely used services can be greatly improved by utilizing new types of data. Many modern fleet vehicles include rich instrumentation such as GPS receivers, sensors to periodically measure sub-system properties [32, 31], and auxiliary computing, storage and communication devices to log and transfer accumulated datasets.

Engine measurement datasets may be used to study the impacts of the environment (e.g., elevation changes, weather), vehicles (e.g., weight, engine size, energy- source), traffic management systems (e.g., traffic light tim- ing policies), and driver behaviors (e.g., gentle acceleration or braking) on fuel savings and GHG emissions. These datasets may include a time-series of attributes such as vehicle location, fuel levels, vehicle speed, odometer values, engine speed in revolutions per minute (RPM), engine load, emissions of greenhouse gases (e.g., CO2 and NOX), etc. Fuel efficiency can be estimated from fuel levels and distance traveled as well as engine idling from engine RPM. These attributes may be compared with geographic contexts such as elevation changes and traffic signal patterns to improve understanding of fuel efficiency and GHG emission. Recent study by the Oak Ridge National Laboratory [8] shows that the heavy truck fuel consumption changes drastically with elevation slope changes. However, these datasets can grow big. For example, measurements of 10 engine variables, once a minute, over the 100 million US vehicles in existence, may have 10^{14} data-items per year. Utilizing such datasets require the development of novel algorithms that exploit modern computing architectures and computational middleware. More details on new challenges posed by big spatial data on mobile applications can be found in [52].

4. CONCLUSIONS

Big spatiotemporal data, though opening up new applications, posing several challenges. New approaches are required to overcome both computational and I/O challenges, and new models that explicitly model spatial and temporal constraints efficiently. Further research is required in the area of compression and sampling. Especially there is a great need for integrating spatial data mining workflows with modern computing infrastructure like cloud computing, in situ [36], data spaces [19], and the like.



(a) Baseball Field (b) Football Field

(c) Building + Car (d) Building + Swimming Pool

Figure 4: Example images sharing similar objects (e.g., grass, buildings, roads, cars, water) but entirely different global labels

5. ACKNOWLEDGMENTS

We would like to thank all our collaborators, especially all the co-authors on various papers cited here. 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.

6. ADDITIONAL AUTHORS

Additional authors: Budhendra Bhaduri (Oak Ridge National Laboratory, bhaduribl@ornl.gov) and Arie Croitoru (George Mason University, acroitor@gmu.edu).

7. REFERENCES

[1] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Ng. Discriminative learning of markov random fields for segmentation of 3d scan data. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 169–176, Washington, DC, USA, 2005. IEEE Computer Society.

[2] L. Anselin. *Spatial Econometrics: methods and models.* Kluwer, Dordrecht, Netherlands, 1988.

[3] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistical Society*, 36:192–236, 1974.

[4] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statistical Soc.*, (48):259–302, 1986.

[5] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report, University of Berkeley, ICSI-TR-97-021, 1997., 1997.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts . *International Conference on Computer Vision*, September 1999.

[7] R. Brittaine and N. Lutaladio. Jatropha: A samllholder bioenergy crop. the potential for pro-poor development. *Integrated Crop Management*, 8:1–114, 2010.

[8] G. Capps, O. Franzese, B. Knee, M. Lascurain, and P. Otaduy. Class-8 heavy truck duty cycle project final report. ORNL/TM-2008/122, 2008.

[9] M. Celik, B. Kazar, S. Shekhar, D. Boley, and D. Lilja. Northstar: A parameter estimation method for the spatial autoregression model. AHPCRC Technical Report No: 2005-001, 2007.

[10] V. Chandola and R. R. Vatsavai. Scalable time series change detection for biomass monitoring using gaussian process. In *NASA Conference on Intelligent Data Understanding (CIDU)*, pages 69–82, 2010.

[11] V. Chandola and R. R. Vatsavai. A gaussian process based online change detection algorithm for monitoring periodic time series. In *SIAM Data Mining (SDM)*, 2011.

[12] V. Chandola and R. R. Vatsavai. A scalable gaussian process analysis algorithm for biomass monitoring. *Statistical Analysis and Data Mining*, 4(4):430–445, 2011.

[13] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. R. Ganguly. Sparse group lasso: Consistency and climate applications. In *SDM*, pages 47–58, 2012.

[14] P. Chou, P. Cooper, M. J. Swain, C. Brown, and L. Wixson. Probabilistic network inference for cooperative high and low levell vision. In *In Markov Random Field, Theory and Applicaitons*. Academic Press, New York, 1993.

[15] N. Cressie. *Statistics for Spatial Data (Revised Edition)*. Wiley, New York, 1993.

[16] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski. Earthquake: Twitter as a distributed sensor system. *Transactions in GIS (in press)*, 0(0), 2012.

[17] D. Das, E. Kodra, Z. Obradovic, and A. R. Ganguly. Mining extremes: Severe rainfall and climate change. In *ECAI*, pages 899–900, 2012.

[18] H. Derin and H. Elliott. Modeling and segmentation of noisy and textured images using Gibbs random fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, (9):39–55, 1987.

[19] C. Docan, M. Parashar, and S. Klasky. Dataspaces: an interaction and coordination framework for coupled simulation workflows. In *HPDC*, pages 25–36, 2010.

[20] S. Fritz, I. McCallum, C. Schill, C. Perger, R. Grillmayer, F. Achard, F. Kraxner, and M. Obersteiner. Geo-wiki.org: The use of crowdsourcing to improve global land cover. *Remote Sensing*, 1(3):345–354, 2009.

[21] A. R. Ganguly and K. Steinhaeuser. Data mining for climate change and impacts. In *ICDM Workshops*, pages 385–394, 2008.

[22] A. R. Ganguly, K. Steinhaeuser, D. J. Erickson, M. Branstetter, E. S. Parish, N. Singh, J. B. Drake, and L. Buja. Higher trends but larger uncertainty and geographic variability in 21st century temperature and heat waves. *Proceedings of the National Academy of Sciences*, 106(37):15555–15559, 2009.

[23] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[24] M. F. Goodchild. *Citizens as Sensors: The World of Volunteered Geography*, pages 370–378. John Wiley and Sons, Ltd, 2011.

[25] J. Graesser, A. Cheriyadat, R. R. Vatsavai, V. Chandola, J. Long, and E. Bright. Image based characterization of formal and informal neighborhoods in an urban landscape. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(4):1164 – 1176, August 2012.

[26] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[27] Y. Jhung and P. H. Swain. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 34(1):67–75, 1996.

[28] G. Jun, R. R. Vatsavai, and J. Ghosh. Spatially adaptive classification and active learning of multispectral data with gaussian processes. In *ICDM Workshops: Spatial and Spatiotemporal Data Mining (SSTDM)*, pages 597–603, 2009.

[29] C. O. Justice, E. Vermote, J. R. Townshend, R. Defries, D. P. Roy, D. K. Hall, V. V. Salomonson, J. L. Privette, G. Riggs, A. Strahler, W. Lucht, R. B. Myneni, Y. Knyazikhin, S. W. Running, S. W. Steve W. Nemani, Z. Wan, A. R. Huete, W. van Leeuwen, R. E. Wolfe, L. Giglio, J.-P. Muller, P. Lewis, and M. J. Barnsley. The moderate resolution imagin spectrradiometer (modis): Land remote sensing for global chang research. *IEEE Transactions on Geosciences and Remote Sensing*, 36:1228–1249, 1998.

[30] S.-C. Kao and A. R. Ganguly. Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios. *J. Geophys. Res.*, 116(D16119), 2011.

[31] H. Kargupta, J. Gama, and W. Fan. The next generation of transportation systems,greenhouse emissions, and data mining. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1209–1212, New York, NY, USA, 2010. ACM.

[32] H. Kargupta, V. Puttagunta, M. Klein, and K. Sarkar. On-board vehicle data stream monitoring using mine-fleet and fast resource constrained monitoring of correlation matrices. *New Gen. Comput.*, 25(1):5–32, Jan. 2007.

[33] J. Kawale, S. Liess, A. Kumar, M. Steinbach, A. Ganguly, N. Samatova, F. Semazzi, P. Snyder, , and V. Kumar. Data-guided discovery of climate dipoles in observations and models. In *NASA Conference on Intelligent Data Understanding (CIDU)*, pages 1–15, 2011.

[34] B. Kazar, S. Shekhar, D. Lilja, R. Vatsavai, and R. Pace. Comparing exact and approximate spatial auto-regression model solutions for spatial data analysis. In *Third International Conference on Geographic Information Science (GIScience2004)*. LNCS, Springer, October 2004.

[35] S. Khan, A. Ganguly, S. Bandyopadhyay, S. Saigal, D. Erickson, V. Protopopescu, and G. Ostrouchov. Non-linear statistics reveals stronger ties between enso and the tropical hydrological cycle. *Geophysical Research Letters*, 33(L24402):6, 2006.

[36] S. Klasky and et. al. In situ data processing for extreme-scale computing. In *SicDAC*, page 16, 2011.

[37] J. LeSage. Bayesian estimation of spatial autoregressive models. *International Regional Science Review*, (20):113–129, 1997.

[38] J. LeSage. Regression Analysis of Spatial data. *The Journal of Regional Analysis and Policy (Publisher: Mid-Continent Regional Science Association and UNL College of Business Administration)*, 27(2):83–94, 1997.

[39] J. P. LeSage and R. Pace. Spatial dependence in data mining. In *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, forthcoming, 2001.

[40] S. Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001.

[41] J. Lovell. Left-hand-turn elimination. New York Times, http://goo.gl/3bkPb„ December 9, 2007.

[42] C. Ma. Spatial autoregression and related spatio-temporal models. *J. Multivarate Analysis*, 88(1):152–162, 2004.

[43] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition and productivity. McKinsey Global Institute, 2011.

[44] G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988.

[45] V. Norris, M. McCahill, and D. Wood. Editorial: The growth of CCTV: a global perspective on the international diffusion of video surveillance in publicly accessible space. *Surveillance and Society*, 2(2/3):110–135, 2004.

[46] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling. Climate data challenges in the 21st century. *Science*, 331(6018):700–702, 2011.

[47] R. Pace and R. Barry. Quick Computation of Regressions with a Spatially Autoregressive Dependent Variable. *Geographic Analysis*, 1997.

[48] R. Pace and R. Barry. Sparse spatial autoregressions. *Statistics and Probability Letters (Publisher: Elsevier Science)*, (33):291–297, 1997.

[49] C. Rasmussen and C. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006.

[50] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704, New York, NY, USA, 2011. ACM.

[51] F. Sahito, A. Latif, and W. Slany. Weaving twitter stream into linked data a proof of concept framework. In *7th International Conference on Emerging Technologies (ICET)*, pages 1–6, 2011.

[52] S. Shekhar, V. Gunturi, M. R. Evans, and K. Yang. Spatial big-data challenges intersecting mobility and cloud computing. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '12, pages 1–6, New York, NY, USA, 2012. ACM.

[53] S. Shekhar, P. Schrater, R. Vatsavai, W. Wu, and S. Chawla. Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transaction on Multimedia*, 4(2):174–188, 2002.

[54] A. H. Solberg, T. Taxt, and A. K. Jain. A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transaction on Geoscience and Remote Sensing*, 34(1):100–113, 1996.

[55] A. Stefanidis, A. Crooks, and J. Radzikowski. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, pages 1–20, 2011.

[56] K. Steinhaeuser, A. Ganguly, and N. Chawla. Multivariate and multiscale dependence in the global climate system revealed through complex networks. *Climate Dynamics*, 39:889–895, 2012.

[57] R. R. Vatsavai. Biomon: a google earth based continuous biomass monitoring system. In *17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems (ACM-GIS)*, pages 536–537, New York, NY, USA, 2009. ACM.

[58] R. R. Vatsavai. Stpminer: a highperformance spatiotemporal pattern mining toolbox. In *Proceedings of the 2nd international workshop on Petascal data analytics: challenges and opportunities*, PDAC '11, pages 29–34, New York, NY, USA, 2011. ACM.

[59] R. R. Vatsavai, A. Cheriyadat, and S. S. Gleason. Unsupervised semantic labeling framework for identification of complex facilities in high-resolution remote sensing images. In *ICDM Workshops*, pages 273–280, 2010.

[60] R. R. Vatsavai, S. Shekhar, and T. E. Burk. An efficient spatial semi-supervised learning algorithm. *International Journal of Parallel, Emergent and Distributed Systems*, 22(6):427–437, 2007.

[61] C. E. Warrender and M. F. Augusteijn. Fusion of image classifications using Bayesian techniques with Markov rand fields. *International Journal of Remote Sensing*, 20(10):1987–2002, 1999.

[62] N. Wayant, A. Crooks, A. Stefanidis, A. Croitoru, J. Radzikowski, J. Stahl, and J. Shine. Spatiotemporal clustering of twitter feeds for activity summarization. In *GIScience (short paper)*, 2012.