Mining for Climate Change

Vipin Kumar
Department of Computer Science and Engineering
University of Minnesota
www.cs.umn.edu/~kumar

The world's population is growing steadily and many countries are simultaneously industrializing [12], developments that have been ongoing at varying rates for two centuries but have accelerated over the past several decades [11]. These processes are increasingly straining already scarce natural and food resources, which must scale up to keep pace with growing demand [3]. The consequences of such large-scale changes include tremendous stresses on the environment [2, 4, 5] that would be calamitous at the current rate of change if they are not managed sustainably [6, 7]. As a result, scientists are tasked with providing answers to challenging questions such as: What is the effect of urbanization on regional land use and ecology? How does deforestation affect the net carbon balance? How does increased biofuel production impact crop patterns and food availability? The ability to address these interconnected, societally-relevant environmental concerns requires development of computational methods that can enable monitoring, analysis and understanding of changes in the Earth system, interactions between different processes, and their impacts on factors such as the carbon cycle, hydrology, air quality, and biodiversity.

Advances in Earth observation technologies have led to the acquisition of vast amounts of accurate, timely, and reliable Earth system data. These rich datasets capture multiple facets of the natural processes and human activities that shape the physical landscape and environmental quality of our planet, and thus offer an opportunity to study the nature of global changes. This availability of new datasets, together with progress in computational data analysis methods, can greatly increase our ability to observe and forecast ecosystem changes [1].

While considerable progress has been made by the Earth sciences in analyzing such datasets to understand our planet and how it is changing, challenges posed by the volume and variety of data have prevented global-scale studies that fully exploit the richness of available information. For example, existing methods may work well when applied for a focused region or a narrowly defined task but fail to work in a global context, e.g., monitoring the health of forests around the world. Thus, there is a strong need for a coordinated effort to advance the state of the art in computational algorithms and robust analysis methods. This call to action is well articulated in the Congressionally-mandated US Global Change Research Program Strategic Plan [10].

Earth science datasets pose several unique challenges: some are due to their inherent spatial-temporal nature and others are specific to the domain. In the following, we describe some of the cross-cutting, data-specific challenges.

1. Spatial-Temporal data
   Many Earth science datasets consist of gridded time series for ecosystem and environmental variables, i.e., each series represents an individual co-registered cell in a latitude-longitude grid that covers the entire surface (or a region) of the Earth. Other

datasets, such as population, river flows and water quality databases, correspond to general spatial regions of a certain scale rather than specific geographic locations. The approaches need to be developed to work with these diverse datasets.

2. Noise
   Earth science datasets are also frequently noisy and/or exhibit high uncertainty due to sensor interference. This issue is particularly acute in the case of remote sensors, where atmospheric (clouds and other aerosols) and surface (snow and ice) interference are constantly encountered. This motivates the need for development of algorithms that are robust to presence of noise and uncertainty in data.
3. Temporal variability
   Ecosystem observations tend to have a high degree of temporal variation, which must be considered in data analysis. For example, the observed greenness of a land area might be different from historical observations due to seasonal change or other changes in soil moisture, precipitation, etc. Handling such naturally occurring temporal variations is necessary to avoid detection of spurious patterns.
4. Multi-resolution and multi-scale
   Ecosystem phenomena occur at different scales. For example, global change events such as urbanization, fires and deforestation tend to impact smaller areas than droughts. Different types of data are also available at different resolutions, often determined by the scale of the measured process. A particular focus of the knowledge discovery process needs to be on building a bridge between these disparate scales and develop approaches that can identify patterns at multiple resolutions.
5. Spatial autocorrelation
   Tobler's first law of geography states that "Everything is related to everything else, but near things are more related than distant things" [9]. This concept of spatial dependence should be leveraged while analyzing ecosystem data.
6. Spatial heterogeneity
   This refers to variability of the observed process over space and is illustrated by natural boundaries of wildfires or deforestation due to topographical constraints, growth of cities along a spatial gradient, or preferential land conversion for agricultural intensification near resources such as lakes and cities. This heterogeneity necessitates the study of irregular-shaped regions in our data mining framework.
7. Scalability
   The available global data for Earth system variables are massive in size, and hence the computational analysis methods must be designed to scale to large datasets.

References Cited

[1] J. S. Clark, S.R. Carpenter, M. Barber, S. Collins, A. Dobson, J. A. Foley, D.M. Lodge, M. Pascual, R.P.Jr., W. Pizer, C. Pringle, W.V. Reid, K.A. Rose, O. Sala, W.H. Schlesinger, D.H. Wall, and D. Wear. Ecological forecasts: An emerging imperative. Science, 293(5530):657–660, 2001.

[2] M. Cropper and C. Griffiths. The Interaction of Population Growth and Environmental Quality. The American Economic Review, 84(2):250–254, 1994.

[3] L.T. Evans. Feeding the Ten Billion: Plants and Population Growth. Cambridge University Press, 1998.

[4] J.A. Foley, R. DeFries, G.P. Asner, C. Barford, G. Bonan, S.R. Carpenter, F.S. Chapin, M. T. Coe, G.C. Daily, H.K. Gibbs, J.H. Helkowski, T. Holloway, E.A. Howard, C.J. Kucharik, C. Monfreda, J.A. Patz, I.C. Prentice, N. Ramankutty, and P.K. Snyder. Global Consequences of Land Use. Science, 309 (5734):570–574, 2005.

[5] J.A. Foley, N. Ramankutty, K.A. Brauman, E.S. Cassidy, J.S. Gerber, M. Johnston, N.D. Mueller, C. O'Connell, D.K. Ray, P.C. West, C. Balzer, E.M. Bennett, S.R. Carpenter, J. Hill, C. Monfreda, S. Polasky, J. Rockstrom, J. Sheehan, S. Siebert, D. Tilman, and D.P.M. Zaks. Solutions for a cultivated planet. Nature, 478(7369):337–342, 2011.

[6] F. Krausmann, S. Gingrich, N. Eisenmenger, K.-H. Erb, H. Haberl, and M. Fischer-Kowalski. Growth in global materials use, GDP and population during the 20th century. Ecological Economics, 68(10): 2696–2705, 2009.

[7] H. Strulik and J.Weisdorf. Population, food, and knowledge: a simple unified growth theory. Journal of Economic Growth, 13(3):195–216, 2008.

[8] United States Global Change Research Program. Strategic Plan 2012-2021. http://strategicplancomments.globalchange.gov, 2011.

[9] W. Tober. A computer movie simulating urban growth in the Detroit region. Economic Geography, 46 (2):234–240, 1970.

[10] United States Global Change Research Program. Strategic Plan 2012-2021. http://strategicplancomments.globalchange.gov, 2011.

[11] U.S. Census Bureau. International Data Base – World Population Summary. http://www.census.gov/population/international/data/idb/worldpopinfo.php.

[12] H. Zhou. Population Growth and Industrialization. Economic Inquiry, 47(2):249–265, 2008.