

*Using Land Surface Phenology for Spatio-temporal Mining of Image Time Series: A Manifesto*

Geoffrey M. Henebry, Ph.D., C.S.E.

& Kirsten M. de Beurs, Ph.D.

*Geoffrey.Henebry@sdstate.edu*

*Kirsten.deBeurs@sdstate.edu*

Geographic Information Science Center of Excellence (GIScCE)

South Dakota State University, Brookings, SD 57007 USA <http://globalmonitoring.sdstate.edu>

**PROLEGOMENA:** We are in an age of intensive earth observation. Critically needed are tools that will enable efficient and accurate characterization of land surface dynamics by analyzing and extracting the spatio-temporal patterns contained in image time series. As domain scientists, rather than computational scientists, we approach the data with some ideas about what constitutes *interestingness* in our data. One recurrent *interesting* theme is the distinction between change from a baseline and variation about a baseline. Baseline characterization is a foundational step in scientifically informed data mining. We seek first to incorporate our domain knowledge in the form of *expectations* of land surface dynamics, so that we can subsequently identify significant deviations from those expectations (de Beurs and Henebry 2004). Indeed, the motivation of extending occasional observation into intensive monitoring is to detect change, quantify disturbance, and enable prediction. The first NASA workshop on earth science data mining identified anomaly detection as a key characteristic of scientific data mining (Behnke et al. 2000).

Data mining approaches were initially focused largely on analysis of business transaction datastreams and governmental databases. Fayyad (1998) divided general data mining techniques into five classes: (1) predictive modeling, which constitutes regression when the data are numeric and classification when categorical; (2) clustering or segmentation of data; (3) summarization techniques which can yield association rules; (4) dependency modeling which seeks latent causal structures; and (5) change and deviation detection, which are typically used with data that are ordered in some manner.

Earth observation datastreams differ in kind from transaction data and periodic statistical tabulations. Shekhar et al. (2004) identified four characteristics of geospatial data that inhibit the application of standard data mining algorithms: (1) rich data types; (2) implicit spatial relationships among variables; (3) observations that are not independent; and (4) spatial autocorrelation among features. Much earth observation data are represented as ratio- or interval-scaled variables rather than categorical variables and, accordingly, the suite of statistical procedures available for data summarization, discrimination, and inference are more powerful.

Association rule discovery has yet to be widely applied to earth observation datastreams (but see Tadesse et al. 2005), probably because most users of these data expect to use their domain expertise to *inform* the data with particular models that can retrieve biogeophysical variables or patterns of interest (e.g., Stolorz et al. 1995; Leen-Kiat and Tsatsoulis 1999). There are relatively few examples of spatio-temporal data mining of biogeophysical data (cf. Roddick and Spiliopoulou 1999; Viña and Henebry 2005) and a recent survey of the clustering techniques for time series revealed relatively little effort applied to earth science data (Liao 2005).

We see two significant issues in spatio-temporal data mining: the selection of appropriate units of analysis and the handling of the structuring effects of autocorrelation. What are the appropriate units of analysis for time series of images that portray variations in electromagnetic

radiation within the context of spatial and temporal coordinates? We assert that georeferenced and temporally located individual pixels are *not* appropriate. Fisher (1997) urges that the pixel is “a snare and a delusion” for it does not constitute a proper geographic object (the mixed-pixel problem) and its ill-defined status often hinders analysis. This warning applies both to imagery *per se* and to its representation and manipulation within GIS (Cracknell 1998). Furthermore, in ecological remote sensing, the pictures themselves are *not* the endpoint of scientific analysis; rather, what is of scientific interest is the dynamic of pattern and process that the pictures portray. Consider the analogy of sparse sampling of individual frames or even frame sequences from a movie. One level of analysis could aim at reconstructing motion from these data, but a more sophisticated analysis could aim at reconstructing the plot. Intelligent, informed knowledge discovery in scientific databases must aim at the latter objective – reconstructing plots, comparing plots, identifying unusual plots as well as interesting deviations from typical plots.

One ecological plot relevant to global change is the seasonality of vegetation growth in temperate climates or, in terms more germane to NASA’s mission, land surface phenology (LSP), which we define as the spatio-temporal patterns of the vegetated land surface as observed by synoptic sensors at spatial resolutions and extents relevant to meteorological processes in the atmospheric boundary layer. We can observe LSP from orbital platforms by sensing *reflected solar* radiation (visible to middle infrared), *emitted terrestrial* radiation (middle infrared through thermal infrared and microwaves), and *backscattered anthropogenic* radiation (radar, lidar).

The process of observation in ecological remote sensing is a more subtle issue than it may first appear. There is the general problem of observability in a strictly technical sense: Is it possible to sample adequately the phenomena of interest? Given the loosely coupled and contingent nature of ecological relationships, this question must be addressed at multiple scales (Allen and Hoekstra 1992), but multi-scale sensing is rarely practiced. Furthermore, the techniques commonly used to represent and manipulate data carry strong assumptions about what constitutes the units of analysis. The traditional units of analysis in remote sensing have been pixels or spectra. Yet, neither is sufficient for measuring spatio-temporal phenomena.

In considering the future of spatial analysis and GIS, Openshaw (1994) argued for a “concepts-rich approach to spatial analysis, theory generation, and scientific discovery in GIS using massively parallel computing.” He diagnosed a source of malaise that continues to affect the spatial analysis community and then points to a remedy:

Pattern searching is not the same as hypothesis testing because there is no relevant null hypothesis. This point was lost on the original quantitative geographers [during the 1970’s]. ... [They] failed to develop a statistical theory of spatial analysis as distinct from providing examples of statistical methods being applied to spatial data in search for largely aspatial patterns. **The danger now is that the same mistake will be repeated 20 years later in the GIS era by a failure to appreciate that spatial patterns are themselves geographic objects that can be recognized and extracted from spatial databases.** [Emphases added.]

The key notion here is that spatial and, by extension, spatio-temporal patterns are *observable entities* and *appropriate units of analysis*. Here is the lever by which to build a theoretical framework for spatio-temporal data mining for image time series. To date, theory development for spatial-temporal analysis has been hampered by lack of a suitable framework for identification and quantification of spatio-temporal patterns. Numerous metrics have been proposed for quantifying spatial properties of image data; however, scant attention has been paid to the effective use of these metrics for capturing or summarizing spatio-temporal dynamics.

Openshaw's critique also points to the problem of baseline models: "...because there is no relevant null hypothesis." The testing of null hypotheses is one particular form of using neutral models to compare and contrast phenomena. Neutral models are touchstones. They serve a crucial role in scientific investigation by providing archetypes of expectation that guide the development of theory, the design of experiments, and the collection, analysis, and interpretation of data. The most powerful inferential tools in traditional probability theory rely upon the concept of zero-dimensional randomness and its formal model, the Gaussian probability distribution function. Similarly, one-dimensional randomness and its formal model, white noise, provide the touchstone for time series analysis. Various spatially random patterns and processes, such as doubly stochastic Poisson processes, self-avoiding random walks, percolation theory, and conditional and simultaneous spatial autoregressive models provide neutral models for two-dimensional data. With the discovery of fractal geometry and the emergence of complexity theory, new neutral models have become useful to characterize distributed-disordered systems: fractional Brownian motion, Ising and Potts models, Levy flights, self-organized criticality, etc.

Notice, however, in this litany of neutral models that abiotic randomness motivates each. This points to a fundamental problem in the use of such neutral models for investigation of biospheric dynamics: the biotic world is not random but, as our ecological understanding demonstrates, is knowable albeit *truly* complex. Many sciences must indirectly observe the responses of "their" dynamical systems to various stimuli, either intentional or coincidental. The problem of inferring process from pattern arises from many-to-one mappings in the absence of domain-specific models to *inform* the inference.

Spatio-temporal observations typically enfold measurements across various phenomenal scales leading to datastreams that exhibit high dimensionality and, from the naïve perspective, many degrees of freedom; however, the interplay of the observed system and the observing process constrains—via autocorrelation—the effective degrees of freedom on possible dynamical expressions. *The "curse of dimensionality" so often bemoaned in data mining literature can be exorcised with **informed** knowledge discovery.* Again, as the dimensionality of the data increases, so does the degree of parameterization in the interaction of pattern and process. In living systems, greater dimensionality yields not only more degrees of freedom but additional degrees of constraint. Myriad new forms of interaction emerge with each additional dimension, but so does structural or functional redundancy. Consider, for instance, about the explosive growth of the symmetry group of an  $n$ -dimensional hypercube as  $n$  increases. Further, many more random neutral models are possible for spatio-temporal dynamics than spatial dynamics because of lagged and periodic effects in space and/or time. How to choose among them? More importantly, how to discover/construct a neutral model that is not anchored to randomness but enables hypothetical articulation and evaluation of complex baseline spatio-temporal patterns? This is a critical issue facing the development of any future environmental monitoring system.

**AN EXAMPLE:** In a recently concluded NASA LCLUC project, we investigated whether the changes in the agricultural sector consequent to the collapse of the Soviet Union had led to changes in land cover and/or land use that would be sufficiently widespread to be observable at spatio-temporal scales that could affect exchanges of water and energy between the land surface and the atmospheric boundary layer. We decided to focus our analysis on the onset of spring because the widespread commencement of vegetation growth causes substantial shifts in the surface energy balance. To model the spring green-up we used two freely available and widely used time series: the Normalized Difference Vegetation Index (NDVI) from the Pathfinder AVHRR Land (PAL) dataset (James and Kalluri 1994) and the near-surface air temperature from

the NCEP/NCAR Reanalysis dataset (Kalnay et al. 1996; Kistler et al. 2001). There are spatio-temporal scale differences in these data. The PAL NDVI is a 10-day maximum value composite at 8km resolution; whereas, the NCEP/NCAR daily temperature extrema are on a 2° global grid. From the latter data we calculated accumulated growing degree-days (AGDD), which is a kind of thermal time easily calculated from daily temperature series. We modeled LSP by linking NDVI to AGDD using two different forms of the relationship: a linear quadratic model that worked well for herbaceous vegetation found in croplands and grasslands (de Beurs and Henebry 2005a) and a nonlinear spherical model that captured well the initial green-up and plateau exhibited by ecoregions dominated by woody vegetation (de Beurs and Henebry 2005b). Both models fit well in some regions and neither does in others. Both models have three parameters that yield four ecologically interpretable metrics that can be mapped annually: the NDVI at the onset of the observing season, the seasonal peak NDVI, the quantity of AGDD needed to reach the peak, and the seasonal dynamic range of NDVI. Model goodness of fit (adjusted  $r^2$ ) provides another important probe into land surface dynamics. We used these LSP models as biometeorological filters on the two image time series to reveal significant shifts in spring's greening in the wake of the Soviet Union. We are currently investigating linkages between modeled LSPs and climate modes (*e.g.*, NAO/AO) in northern Eurasia as part of NEESPI (Figures 1-3). In challenging the image time series with specific functional models informed by our ecological understanding, we enhance our ability to detect where we understand the data and where the models break down.

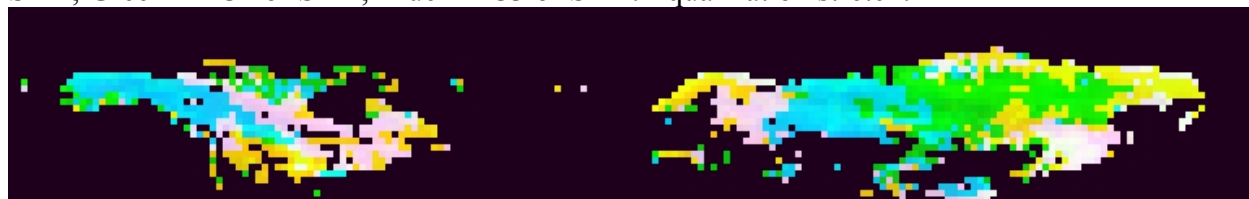
Figure 1: Goodness of fit map of LSP models for the northern hemisphere portrayed using NCEP/NCAR 2°x2° gridcells: first PC of 9 yrs of PAL NDVI data (1985-88 and 1995-99, *cf.* de Beurs and Henebry 2004 for years selected). Black gridcells have  $r^2 < 0.5$ . 2% stretch.



Figure 2: Ecological expectations for LSP models of the northern hemisphere portrayed using NCEP/NCAR 2°x2° gridcells: Red = PC1 of NDVI at start of observing season; Green = PC1 of AGDD to reach peak NDVI; Blue = PC1 of seasonal dynamics range of NDVI. 2% stretch.



Figure 3: Major modes of spatio-temporal variation in seasonal dynamic range (SDR) in LSP models for the northern hemisphere portrayed using NCEP/NCAR 2°x2° gridcells: Red = PC2 of SDR; Green = PC1 of SDR; Blue = PC3 of SDR. Equalization stretch.



## REFERENCES

- Allen, T.F.H. and T. Hoekstra. 1992. Toward a Unified Ecology. New York: Columbia University Press.
- Behnke, J., E. Dobinson, S. Graves, et al. 2000. *Final Report: NASA Workshop on Issues in the Application of Data Mining to Scientific Data*.
- Cracknell, A.P. 1998. Review article. Synergy in remote sensing – what’s in a pixel? *International Journal of Remote Sensing* 19:2025-2047.
- de Beurs, K.M., and G.M. Henebry. 2004. Trend analysis of the Pathfinder AVHRR Land (PAL) NDVI data for the deserts of Central Asia. *IEEE Geoscience and Remote Sensing Letters* 1(4):282-286.
- de Beurs, K.M., and G.M. Henebry. 2005a. A statistical framework for the analysis of long image time series. *International Journal of Remote Sensing* 26(8):1551-1573.
- de Beurs, K.M., and G.M. Henebry. 2005b. Land surface phenology and temperature variation in the IGBP high-latitude transects. *Global Change Biology* 11(5):779-790.
- Fayyad, U. 1998. Mining databases: towards algorithms for knowledge discovery. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 21:39-49.
- Fisher, P. 1997. The pixel: a snare and a delusion. *International Journal of Remote Sensing* 18:679-685.
- James, M.E., and S.N.V. Kalluri. 1994. The Pathfinder AVHRR Land data set - an improved coarse resolution data set for terrestrial monitoring. *International Journal of Remote Sensing* 15: 3347-3363.
- Kalnay, E., M. Kanamitsu, R. Kistler, et al. 1996. The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, 77: 437-471.
- Kistler, R., E. Kalnay, W. Collins, et al. 2001 The NCEP-NCAR 50-Year Reanalysis: monthly mean CD-ROM and documentation. *Bulletin of the American Meteorological Society* 82:247-267.
- Leen-Kiat, S., and C. Tsatsoulis. 1999. Segmentation of satellite imagery of natural resources using data mining. *IEEE Transactions on Geoscience and Remote Sensing* 37:1086-1099.
- Liao, T.W. 2005. Clustering of time series data – a survey. *Pattern Recognition* 38:1857-1874.
- Openshaw, S. 1994. A concepts-rich approach to spatial analysis, theory generation, and scientific discovery in GIS using massively parallel computing. In: (M. Worboys, ed.) Innovations in GIS. London: Taylor and Francis. pp. 123-137.
- Roddick, J.F., and M. Spiliopoulou. 1999. A bibliography of temporal, spatial and spatio-temporal data mining research. *SIGKDD Explorations* 1:34-38.
- Shekhar, S., P. Zhang, Y. Huang, and R. R. Vatsavai. 2004. Trends in spatial data mining. In: (H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, eds.), Data Mining: Next Generation Challenges and Future Directions. Cambridge, MA: MIT/AAAI Press. pp. 357-380.
- Stolorz, P., H. Nakamura, E. Mesorobian, et al. 1995. Fast spatio-temporal data mining of large geophysical datasets. *Proc. First International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Montreal. pp. 300-305.
- Tadesse, T., D.A. Wilhite, M.J. Hayes, et al. 2005. Discovering associations between climatic and oceanic parameters to monitor drought in Nebraska using data-mining techniques. *Journal of Climate* 18:1541-1550.
- Viña, A., and G.M. Henebry. 2005. Spatio-temporal change analysis to identify anomalous variation in the vegetated land surface: ENSO effects in tropical South America. *Geophysical Research Letters* 32:L21402.