# "The (Statistical) Science of Sustainability" A White Paper for the Workshop on Mathematical Challenges for Sustainability (DIMACS Center, Rutgers University November 15-17, 2010)

Noel Cressie (ncressie@stat.osu.edu)
Program in Spatial Statistics and Environmental Statistics
Department of Statistics
The Ohio State University
Columbus OH 43210-1247

November 18, 2010

## Preamble

I take as a basic law of nature that "trees do not grow to the sky." In fact, the tree metaphor is very useful as we try to impose a science on the vaguely defined term, "sustainability." Trees make up forests and forests grow and recede according to many factors. Trees develop from seeds to saplings to mature trees, using nutrients and water in the soil and $CO_2$, $O_2$ and light from the atmosphere to grow. They do not grow to the sky, and they eventually die. Forests do not cover the earth; but they grow, recede, and are potentially sustainable. Trees and forests are a function of their environment. *Homo sapiens* are a function of their environment too, although we sometimes forget.

Achieving sustainability should be closely linked to ensuring the environment will not change precipitously. Ecology is the study of organisms and how they relate; environmental science is the study of the surroundings of those organisms. There is a hope that the functions that relate organisms to their environments are smooth. Chaos theory shows that this is not always the case (e.g., Devaney, 1989). Uncertainty in the functional relationships leads to a theory of random functions (i.e., stochastic processes); see, for example, Karlin and Taylor (1981) and Adler (1981).

Even when a function is known and smooth, an abrupt environmental change will lead to enormous disruption of the life cycle of organisms and their inter-relationships (including trees and humans!). Life, as we know it in 2010, may not be sustainable, and to make it so, we need measures and an accompanying Science of Sustainability (SOS!). Broadly speaking,

we should be trying to understand the behavior of these functions and, for those that have derivatives, we should be trying to calculate their derivatives temporally and spatially.

## Sustainability of What?

In my preamble, I have concentrated on sustainability of life on this planet. But life is complex and there are other planets. Leaving aside a grand plan to find other habitable planets, we should recognize that micro-scale structures beget small-scale structures, which beget large-scale structures, and they can be intricately linked in a feedback relationship; life *is* complex. At the micro scale, food, water, shelter, and reproduction are huge "forcing" terms. *Homo sapiens* also experience happiness, greed, jealousy, etc; they value art and music too, and all of these can be important "forcing" terms. It is tempting to try to put them all into a basket and compute a CSI (Consumer Sustainability Index). This would be a mistake; the structures are highly multivariate and complex.

Laws of nature, psychology, finance, and policy might *contribute to* "laws of sustainability," but we are embarking on a complexity that goes beyond those laws. Moreover, our three-dimensional spatial world, evolving dynamically in one-dimensional time, makes the science of sustainability even more complex.

Mathematically, the complexity I have been describing can be captured by graphs (commonly called networks). There is an embryonic discipline of Complex Networks, populated by physicists, statisticians, computer scientists, epidemiologists, mathematicians, etc. The NSF-funded Statistical and Applied Mathematical Sciences Institute (SAMSI) has a program devoted to it in 2010-2011 (www.samsi.info/programs/2010cnprogram.shtml); also see, for example, the books by Kolaczyk (2009) and Newman (2010). Complex networks are *models* of how "the world works," but they are extendable to allow for more complexity or more variables. That strength is also a weakness; network sizes and complexities can clearly grow exponentially. But, it does offer a paradigm to understand growth and its counterpart, recession. In fact, sustainability will, by definition, require *both* growth and recession, in different sectors, in different regions, and at different times.

To build, measure, and assimilate a complex network is a worthy endeavor, but destined to fail. An analogy would be to try to track every gas molecule in the atmosphere over time. One way to view this is whether we take a "field view" or an "object view" of the part of the world we are trying to model.

The *object* view of the world sees individual objects, located in a spatial domain and interacting through time with each other, often as a function of their "distance" apart. Thus, a household and its characteristics make up a unit of interest to census enumerators. This microdatum is typically unavailable to social scientists, for confidentiality reasons. Consequently, the census data that are released are typically the *number of objects* in small areas, but not their locations. That is, a set of count data from small areas is released, which is simply an aggregated version of the object view of the world. The geographical extent (i.e., spatial support) of a small area can be stored in a Geographical Information System (GIS) as a polygon. (A GIS is a suite of hardware and software tools that features linked georeferencing in its database management and in its visualization.)

Alternatively, the *field* view of the world loses sight of the objects and potentially has

a (multivariate) datum at every spatial location in the domain of interest. Building on the census-enumeration example discussed above, we can define a *field* as the object density, in units of number per unit area, at any location. This is purely a mathematical construct because, at a given location, either there is an object present or there is not. Such a density can be estimated from a moving window, such that at any location the estimated density is the number of objects per unit area in the window at that location.

Sometimes the field view is the result of an aggregation of the object view, as for population-density data. Other times, the field view is all that is of interest, such as for rainfall data where there is typically no interest in the individual raindrops. Again, a GIS is a convenient way to store data for a field, along with the spatial support to which a datum refers.

A useful way to visualize the difference between the object view and the field view is to imagine yourself in a helicopter taking off from a clearing in a field of corn. As the helicopter ascends, at some point it is no longer interesting to think of objects (e.g., corn plants), but rather to think of a field, literally and statistically (e.g., in units of bushels per acre). Then the temporal aspect is captured through the field's dynamical evolution during the growing season.

Applying these ideas to complex networks means that we should move away from their mathematical building blocks (vertices and edges) and study instead identifiable "objects" made up of those building blocks and/or study local densities ("fields") within the network. There is a strong analogy in image analysis, where the study of an image through its pixels has its limitations. For image reconstruction it might be appropriate but, for target tracking, a (random) set-based approach is considerably more powerful (e.g., Grenander and Miller, 2007). The solution depends on the question.

# Uncertainty

Our world is uncertain, our attempts to explain the world (science) are uncertain, and our measurements of our (uncertain) world are uncertain. To build a Science of Sustainability, it is important to recognize and quantify the uncertainty in measuring, modeling, and predicting; this is true for all of science and its myriad fields.

Hierarchical statistical modeling represents a way to express uncertainties through well defined levels of conditional probabilities. At the top level is the *data model*, which expresses the distribution of the data given a hidden process. At the level directly underneath the data model is the *process model*, which models scientific uncertainty in the hidden ("true") process through a probability distribution of the phenomenon of interest. It is quite possible that the process model is itself made up of sub-models whose uncertainties are also expressed at sub-levels through conditional probabilities. In a sense, the whole approach is a sort of analysis-of-variance decomposition that is more general than the usual additive decomposition given in standard statistics textbooks. The result is a *hierarchical model* (HM).

The components of an HM are *conditional probability distributions* that, when multiplied together, yield the joint probability distribution of all quantities in the model. The quantities in which we are interested could be as simple as random variables and as complicated as space-time stochastic processes of random objects.

Of course, all the conditional probability distributions specified in the HM typically depend on unknown parameters. If a lower level (underneath the data model and the process model) is established by specifying the joint probability distribution of all the unknown parameters, then the HM is called a *Bayesian Hierarachical Model* (BHM). This probability model at the lowest level, which we call the *parameter model*, completes the sequence: data model (top level) followed by process model (second level) followed by parameter model (bottom level). An alternative approach to specifying the parameter model is to *estimate* the parameters using the data. This might be called an *Empirical Hierarchical Model* (EHM), although historically it has often been called an empirical-Bayesian model. The HM approach is explored for complex spatio-temporal models in Cressie and Wikle (2011).

The HM paradigm enables a coherent use of all data and allows inference on parts where there are *no* data at all! Scientific relationships incorporated into the process and parameter models can mitigate the paucity of data. Further, there is a self-correcting mechanism in hierarchical statistical modeling, namely, when there is little known about the scientific relationships or there are poor-quality or few data available, then inferences have very low precision. That is, a signal in the process may be present, but if scientific knowledge or the data are limited, the HM approach will *not let us* discover it.

Looking at this from another angle, the best scientists collect the best data to build the best (conditional-probability) models to make the most precise inferences in the shortest amount of time. In reality, compromises at every stage may be needed, and we could add that the best scientists make the best compromises!

# References


Adler, R. (1981). *The Geometry of Random Fields*. Wiley, New York, NY.

Cressie, N. and Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.

Devaney, R.L. (1989). *Introduction to Chaotic Dynamical Systems*, 2nd edn. Benjamin-Cummings, Menlo Park, CA.

Grenander, U. and Miller, M. (2007). *Pattern Theory: From Representation to Inference*. Oxford University Press, Oxford, UK.

Karlin, S. and Taylor, H.M. (1981). *A Second Course in Stochastic Processes*. Academic Press, San Diego, CA.

Kolaczyk, E.D. (2009). *Statistical Analysis of Network Data*. Springer, New York, NY.

Newman, M.E.J. (2010). *Networks: An Introduction*. Oxford University Press, Oxford, UK.