Gaps and Opportunities for Machine Learning and Geostatistics in the Natural Sciences

Michael L. Stein

University of Chicago

August, 2018
Some background on geostatistics

Arose out of mining applications in the 1950s and 1960s (Krige, Matheron).
  ▶ At first developed largely outside of the mainstream statistics community.
  ▶ Main problem of interest was spatial prediction, especially of area averages, based on limited observations.
  ▶ Focus on second-order methods (means, variances and covariances).
  ▶ Semivariogram, $\gamma(h) = \frac{1}{2}E\{Z(x) - Z(x + h)\}^2$, used to model spatial dependence.

Second-order methods avoids assumptions of probability distributions.
  ▶ However, I would argue that classical geostatistics really only makes sense for processes that are close to Gaussian.

What are most important properties of Gaussian processes when interest is prediction?
Want to know conditional distributions of unknowns given the observations. For Gaussian processes, important properties, from most to least:

- Conditional expectations are linear in the observations.
- Conditional variances are independent of observed values.
- Conditional distributions are Gaussian.

For spatial processes with strong local dependence (most environmental processes), the fact that marginal distributions are Gaussian hardly matters. Even if one is just interested in an accurate stochastic model (high likelihood for data)

- Conditional distributions of observations given neighboring observations is what matters most.

So what does this tell us about most important problems in statistical models for space-time data?
Three big issues, in order of **decreasing** importance:

- Models
- Diagnostics
- Computation

Growing size of environmental datasets may lead us to focus too much on computation.

Paraphrasing Tukey, approximate computations on an appropriate model is more useful than exact computations on an ill-suited model.

Without good diagnostics, development of better models will be problematic.
Even just for Gaussian processes, much to do for space-time covariance functions:

- Space-time interactions
- Nonstationarities
- Seasonal/diurnal cycles
- Distinguishing long-term trends from stochastic variation
- Vertical dimension
- Multivariate dependencies

GOES geostationary satellites as source of space-time data that is high-resolution in both space and time.

Some issues for non-Gaussian processes

- Limited utility of processes that are conditionally independent given a hidden Gaussian process.
- Need to focus on conditional properties.
  - Joint conditional distribution of daily precipitation at many unobserved sites given rain gauge data?
- For Gaussian processes, spectral methods useful for modeling and diagnostics. Right analogs (e.g., wavelets) for non-Gaussian processes?
- Getting appropriate science into models.
How much data?

In my experience, there is no such thing as “enough” data. For complex environmental processes, the more data that are available, the more one can see new features in the process.

Some processes now observed so densely in space-time that interpolation is no longer needed? Still need statistics.

- More data will never solve the forecasting problem.
- Even in fields with highly informative physical models (e.g., meteorology), statistical methods needed at the least for appropriate UQ.
  - Covariance matrices for data assimilation.
- For many complex processes (e.g., precipitation) should include intrinsic stochastic elements to capture variation realistically.
- Good stochastic model informative about process (e.g., nonstationarities on global scale).

Brief comment on ML for environmental data:
- Methods for analyzing space-time data should make explicit use of space-time information.
  - Principal components does not.
  - Classical geostatistics does.