A Case Study:

High Resolution Soil Water from Regional Databases and Satellite Images

Robin D. Morris, Vadim N. Smelyanskiy

in collaboration with Joseph Coughlan, Jennifer Dunjan

Combating uncertainty:

- Before you can combat uncertainty, you have to honestly quantify the uncertainty present.
- Representing uncertainty in terms of probabilities is the natural framework, and allows a huge body of statistical tools to be readily applied.
- It also corresponds to what we do - "quantified common sense".
- "The theory of probabilities is based on certain axioms which seem plausible ..." - Laplace
- Tools now exist that allow for computation in complex statistical models

Case study:

- Soil water is often a limiting factor in plant growth.
- Ecologists are therefore interested in soil water over large areas.
- Directly measuring soil water is difficult, involving extensive fieldwork and laboratory work.
- Satellite data can be used to infer plant growth.
- Under certain circumstances, plant growth can be linked to soil water.

Inputs:

- Satellite observations
  - AVIRIS data and near infrared bands at 0.95 um wavelength
- STATSGO database
  - Soil geographic database
  - Provides soil water information on soil water from an unknown (and most probably unseen) and irregular sampling of the region
- Task is to relate AVIRIS data to soil water and combine with the STATSGO information

Probabilistic formulation:

- Start with the distribution of what we do know (AVIRIS data), conditional on everything we don't know (NDVI, LAI, AWC)
  \[ p(\text{NDVI}, \text{LAI}, \text{AWC}) | p(\text{AVIRIS}) \]
- Invert using Bayes' Theorem
  \[ p(\text{NDVI}, \text{LAI}, \text{AWC}) | p(\text{AVIRIS}) = p(\text{NDVI}, \text{LAI}, \text{AWC}) | p(\text{AVIRIS}) \]
- Look at the conditional independence structure of the problem
  - This is shown on the previous diagram (repeated below)
Probabilistic formulation (cont):

- \( p(\text{aviris \mid ndvi, lai, awe}) \cdot p(\text{ndvi, lai, awe}) \)
  - general, avoids data in measurement at end of week
- \( p(\text{aviris \mid ndvi}) \cdot p(\text{ndvi, lai, awe}) \)
  - new test, fai data is measurement at end of week
- \( p(\text{aviris \mid ndvi}) \cdot p(\text{ndvi, lai, awe}) \cdot p(\text{la, awe}) \)
  - general test, ndvi is independent of awe
- \( p(\text{aviris \mid ndvi}) \cdot p(\text{ndvi, lai}) \cdot p(\text{la, awe}) \)
  - regular joint distribution, p(\text{la, awe})
- \( p(\text{aviris \mid ndvi}) \cdot p(\text{ndvi}) \cdot p(\text{lai, awe}) \cdot p(\text{awe}) \)
  - look at each term in this last expression and assess the uncertainty involved

Sources of uncertainty:

- \( p(\text{aviris \mid ndvi}) \cdot p(\text{ndvi, lai, awe}) \cdot p(\text{lai, awe}) \cdot p(\text{awe}) \)
  - NDVI is derived directly from AVIRIS read and near infrared bands.
  - NDVI = \( R - N(R - RED) \)
  - compared with other terms, the uncertainty here is negligible.
- \( p(\text{aviris \mid ndvi}) \cdot p(\text{ndvi, lai, awe}) \cdot p(\text{lai, awe}) \cdot p(\text{awe}) \)
  - field data - growing season
  - no data on the satellite sensor
  - airborne resampling

Sources of uncertainty (cont):

- \( p(\text{aviris \mid ndvi}) \cdot p(\text{ndvi, lai, awe}) \cdot p(\text{lai, awe}) \cdot p(\text{awe}) \)
  - more fieldwork, measuring LAI for a number of plots, and also making laboratory measurements of soil water.
  - Nemani & Running (1988) have the data points for this graph:
  - The form of the curve, \( \text{NDVI} = A \log(B \times \text{LAI}) \), comes from ecological theory.
  - But what about the parameters?

Sources of uncertainty (cont):

- \( p(\text{aviris \mid ndvi}) \cdot p(\text{ndvi, lai, awe}) \cdot p(\text{lai, awe}) \cdot p(\text{awe}) \)
  - this is the second source of information - the STATSGO database provides prior information on the distribution of AWC values.
  - map polygons, components, layers
  - for each layer we are given the range and range that was measured in the sampling, and the mean and std dev.
  - for each component we are given the % of the polygon that is the same component.
  - assuming that the sampling misses the tails of the distribution, we form a mixture model where the maximum values are taken as \( \mu \pm 2 \sigma \) for that layer.
Putting it all together:

- $p(\text{awc}, \text{lai} | \text{ndvi}) = p(\text{ndvi} | \text{lai}) p(\text{lai} | \text{awc}) p(\text{awc})$
- what we're really interested in is:
  - $p(\text{awc} | \text{ndvi}) = p(\text{awc}, \text{lai} | \text{ndvi})$
- This can be approximated by sampling from $p(\text{awc}, \text{lai} | \text{ndvi})$ and then ignoring the lai values.

Results:

Oregon NDVI
Polygon OR149

Results (cont):

Distribution of AWC from sampling, excluding the STATSGO prior

Discussion:

- The awc distributions shown are for sampling $p(\text{ndvi} | \text{lai}) p(\text{lai} | \text{awc})$, not including the prior $p(\text{awc})$.
- Compare the distribution from sampling with the STATSGO prior:
- See the uncertainty is still on the level of the components.

Further refinements:

- Improving $p(\text{ndvi} | \text{lai})$ and $p(\text{lai} | \text{awc})$ can be done by incorporating other types of information, e.g.
- tree species,
- age,
- time since last cleanup etc.
- but all of these require more fieldwork.
- Including a spatial prior will reduce variation across the polygons.
Representing distributions using samples

- We wish to represent \( p(x) \), and then to compute expectations
  \[
  E(f(x)) = \int f(x) p(x) dx
  \]
- If we sample the domain of \( x \) uniformly, and compute \( p(x) \) for each \( x \) value, the integral can be approximated by
  \[
  E(f(x)) = \sum_i f(x(i)) p(x(i))
  \]
- However this becomes inefficient if \( x \) is in more than a few dimensions (curse of dimensionality)
- Instead, if we sample from \( p(x) \), is concentrate the samples in the high probability regions, then we can approximate the integral by
  \[
  E(f(x)) = \sum_i f(x(i)) p(x(i))
  \]

Markov chain Monte Carlo:

- MCMC is a method of generating samples from \( p(x) \).
- A Markov chain is a sequence of states, where the probability of the next state depends only on the current state.
- The method constructs a Markov chain that converges to the distribution of interest, \( p(x) \).
- To do this it is necessary to determine the Transition Probabilities.
- The Metropolis Algorithm is the simplest scheme for doing this:
  - Initialize \( x \) randomly
  - Propose a new value \( x' \), where one of the elements of \( x \) is changed by drawing it from a symmetric distribution
  - Accept \( x' \) as the new value with probability
    \[
    p_{new} = \min \left( 1, \frac{p(x')}{p(x)} \right)
    \]
  - otherwise retain the current value, \( x \)
  - store the realizations
- The realizations are an approximate sample from the posterior, from which we can compute quantities of interest (means, variances etc).