A Case Study:
High Resolution Soil Water from Regional Databases and Satellite Images
Robin D. Morris, Vadim N. Smelyanskiy
in collaboration with Joseph Doughan, Jennifer Duban

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".

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 clear near infrared bands at 1.5m resolution
- STATSGO database
  - Wide soil property database
  - provides related information on soil water from an unknown (and model potential) source
  - regular sampling of the input
- 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)
  \[ p(\text{AVIRIS} | \text{NDVI, LAI, AWC}) \]
- Invert using Bayes' Theorem
  \[ p(\text{AVIRIS} | \text{NDVI, LAI, AWC}) = \frac{p(\text{AVIRIS}) p(\text{NDVI, LAI, AWC} | \text{AVIRIS})}{p(\text{NDVI, LAI, AWC})} \]
- Look at the conditional independence structure of the problem
  - this is shown in the probability diagram (repeated below)

\[ \text{satellite data} \rightarrow \text{NDVI} \rightarrow \text{leaf area index} \rightarrow \text{soil water} \]
**Probabilistic formulation (cont):**

Sources of uncertainty:

- **p(\text{aviris} | \text{ndvi, lai, awe}) = p(\text{ndv}, \text{lai, awe})**
  - green algal activity is a non-linear interaction
  - p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai, awe})
  - green algal activity independent of soil water
- **p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai, awe})**
  - green algal activity independent of soil water
- **p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai}) p(\text{lai, awe}) p(\text{awe})**
  - Look at each term in this last expression and assess the uncertainty involved

**Sources of uncertainty (cont):**

- **p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai, awe})**
  - This comes from fieldwork - measuring LAI for a number of plots, and finding the NDVI values for these plots.
  - Derived from AVIRIS images taken at the same time.
  - Running et al. (1989) have the following graph:

![Graph showing the relationship between AVIRIS and field data](image)

But what about the parameters?

- **p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai}) p(\text{lai, awe}) p(\text{awe})**
  - This is the second source of information - the STATSGO database provides prior information on the distribution of AWC values.

**Sources of uncertainty (cont):**

- **p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai}, awe) p(\text{lai, awe})**
  - Statistics from AVIRIS are indirectly related to AWC levels.
  - Compared with the other terms, the uncertainty here is negligible.

If we assume that
\[ \text{ndvi} = A \log(B \times \text{lai}) = e \]
where \( e \sim N(0, \sigma) \)

then we can write
\[ p(A, B, e | \text{data}) = p(\text{data} | A, B, e) p(A, B, e) \]

and realizations from \( p(A, B, e | \text{data}) \) can be used to represent this distribution (more on this later).

**Sources of uncertainty (cont):**

- **p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai}, awe) p(\text{lai, awe}) p(\text{awe})**
  - Cornplined with the other terms, the uncertainty here is negligible.

- **p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai}) p(\text{lai, awe}) p(\text{awe})**
  - Running et al. (1989) have the following graph:

Again, sample from \( p(A, B, e | \text{data}) \)
where \( \text{lai} = A x \text{awe} + B + e \)

\[ p(\text{lai, awe}) = N(A x \text{awe} + B, \sigma) \]

For each layer we are given the top and soil text; this text was converted to the sampling, and the top and soil depth.

For each component we are given the % of the polygon that this component occupies.

Assuming that the sampling misses the tails of the distribution, we form a mixture model where the maximum values are taken as \( \infty \) for that layer.

**Sources of uncertainty (cont):**

- **p(\text{aviris} | \text{ndvi}) p(\text{ndvi, lai, awe}) p(\text{lai, awe}) p(\text{awe})**
  - This comes from fieldwork - measuring LAI for a number of plots, and finding the NDVI values for these plots.
  - Derive from AVIRIS images taken at the same time.
  - Running et al. (1989) have the following graph:

![Graph showing the relationship between AVIRIS and field data](image)
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
- \( p(\text{awc} | \text{ndvi}) = (p(\text{awc} | \text{lai}) | p(\text{lai} | \text{ndvi}) \)

- This can be approximated by sampling from \( p(\text{awc}, \text{lai} | \text{ndvi}) \) and then ignoring the lai values.

Results (cont):

- Distribution of AWC from sampling, excluding the STATSGO prior

Discussion:

- The awc distributions shown are for sampling \( p(\text{ndvi}, \text{lai} | \text{awc}) \), 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 harvest 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, then the integral can be approximated by
  \[ E(f(x)) \approx \sum f(x(i))p(x(i)) \]
- 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
  \[ E(f(x)) \approx \sum f(x(i)) \]

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 or the elements of \( x \) is changed by drawing it from a symmetric distribution
  - Accept \( x' \) as the new value with probability \( p = \min \left( 1, \frac{p(x')}{p(x)} \right) \)
  - 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).