59-128

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 Dungan

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 at bottom only common sense reduced to computation." - Laplace

- Tools now exist that allow for computation in complex statistical models.

Case study:

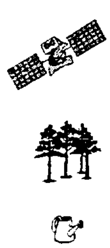
- Soil water is often a limiting factor in plant growth.
- Ecologists are therefore interested in soil water over *large areas*.
- Directly measuring soil water is difficult, involving extensive fieldwork and laboratory work.

- Satellite data can be used to infer plant growth.
- Under certain circumstances, plant growth can be linked to soil water.

Inputs:

- Satellite observations
  - AVIRIS red and near infra-red bands at 1.1km resolution
- STATSGO database
  - state soil geographic database
  - provides limited information on soil water from an unknown (and most probably sparse and irregular) sampling of the region
- Task is to relate AVIRIS data to soil water and combine with the STATSGO information.

- Under certain circumstances, plant growth can be linked to soil Available Water Capacity (AWC).
- Leaf Area Index (LAI) quantifies plant growth.
- NDVI (non-dimensional vegetation index) relates leaf area to the AVIRIS data.
- The Leaf Area Index and NDVI act as a *latent variables* - linking the observations to the quantity of interest.



satellite data ← NDVI ← leaf area index ← soil water

Probabilistic formulation:

- Start with the distribution of what we *do* know (AVIRIS data), conditional on everything we don't know (NDVI, LAI, AWC)

  p(aviris | ndvi, lai, awc)

- Invert using Bayes' Theorem
  - because what we're really interested in is p(ndvi, lai, awc | aviris)
- p(ndvi, lai, awc | aviris) ∝ p(aviris | ndvi, lai, awc) p(ndvi, lai, awc)
- Look at the *conditional independence structure* of the problem
  - this is shown on the previous diagram (repeated below)

satellite data ← NDVI ← leaf area index ← soil water

Probabilistic formulation (cont):

- p(aviris | ndvi, lai, awc) p(ndvi, lai, awc)
  - given ndvi, aviris data is independent of lai and awc
- p(aviris | ndvi) p(ndvi, lai, awc)
  - rewrite the joint distribution p(ndvi, lai, awc)
- p(aviris | ndvi) p(ndvi | lai, awc) p(lai, awc)
  - given leaf area, ndvi is independent of awc
- p(aviris | ndvi) p(ndvi | lai) p(lai, awc)
  - rewrite the joint distribution p(lai,awc)
- p(aviris | ndvi) p(ndvi | lai) p(lai | awc) p(awc)
- Look at each term in this last expression and assess the uncertainty involved

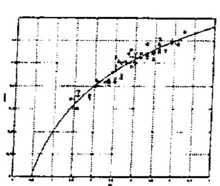
Sources of uncertainty:

- p(aviris | ndvi) p(ndvi | lai) p(lai | awc) p(awc)
- ndvi is derived directly from AVIRIS read and near infra-red bands.
  - ndvi = (IR - RED)/(IR+RED)
- Compared with the other terms, the uncertainty here is negligible.
  - noise on the satellite sensor
  - geometric resampling

Combating Uncertainty With Fusion Workshop | Robin Morris, RIACS, NASA Ames

Sources of uncertainty (cont):

- p(aviris | ndvi) p(ndvi | lai) p(lai | awc) p(awc)
- This comes from fieldwork - measuring LAI for a number of plots, and finding the NDVI values for those plots.
  - Derived from an AVIRIS image taken at the same time.
- Running et al (1989) have the following graph:



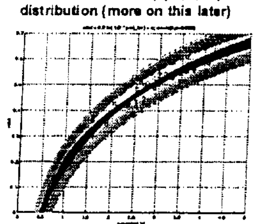
The *form* of the curve,
ndvi = A log(B x lai),
comes from ecological theory.

But what about the parameters?

Combating Uncertainty With Fusion Workshop | Robin Morris, RIACS, NASA Ames

Sources of uncertainty (cont):

If we assume that ndvi = A log(B x lai) + e
where e <- N(0, σ)
then we can write
p(A,B, σ | data) = p(data | A,B, σ) p(A,B, σ)
and *realizations* from p(A,B, σ | data) can be used to represent this distribution (more on this later)



blue lines: curves from (A,B) values

green lines: +/- 2 σ

slicing through this graph we can form
p(ndvi | lai) = N(*A* log(*B*xlai), σ)

where *(A,B, σ) are mean values*

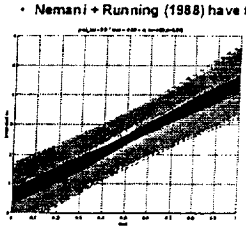
Note: using a single value underestimates the uncertainty in the distribution.

Assuming a known *form* for the relationship is very strong information, and is why the variance does not explode outside the range of the data.

Combating Uncertainty With Fusion Workshop | Robin Morris, RIACS, NASA Ames

Sources of uncertainty (cont):

- p(aviris | ndvi) p(ndvi | lai) p(lai | awc) p(awc)
- 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:



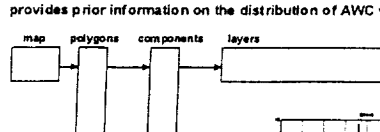
Again, sample from p(A,B, σ | data)
where lai = A x awc + B + e

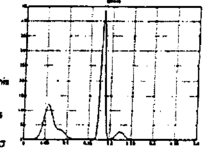
p(lai | awc) = N(*A* x awc + *B*, σ)

Combating Uncertainty With Fusion Workshop | Robin Morris, RIACS, NASA Ames

Sources of uncertainty (cont):

- p(aviris | ndvi) p(ndvi | lai) p(lai | awc) p(awc)
- This is the second source of information - the STATSGO database provides prior information on the distribution of AWC values.





For each layer we are given the min and max awc that was recorded in the sampling, and the max and min depth.

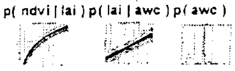
For each component we are given the % of the polygon that this component comprises.

Assuming that the sampling misses the tails of the distribution, we form a mixture model where the max/min values are taken as +/- 2σ for that layer.

Combating Uncertainty With Fusion Workshop | Robin Morris, RIACS, NASA Ames

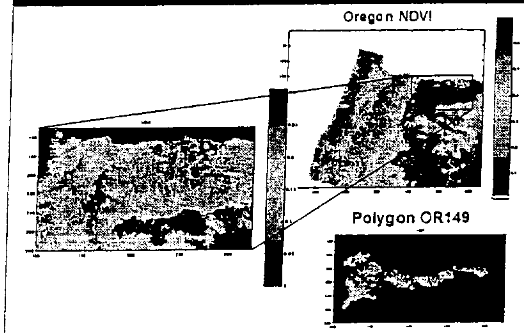
Putting it all together:

- $p(\text{awc}, \text{lai} \mid \text{ndvi}) = p(\text{ndvi} \mid \text{lai})\, p(\text{lai} \mid \text{awc})\, p(\text{awc})$

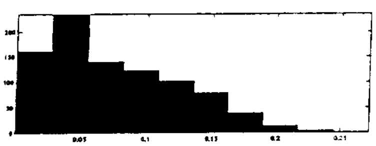


- what we're really interested in is
- $p(\text{awc} \mid \text{ndvi}) = \int_{\text{lai}} p(\text{awc}, \text{lai} \mid \text{ndvi})$
- This can be approximated by sampling from p(awc, lai | ndvi) and then ignoring the lai values.

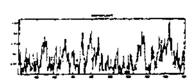
Results:



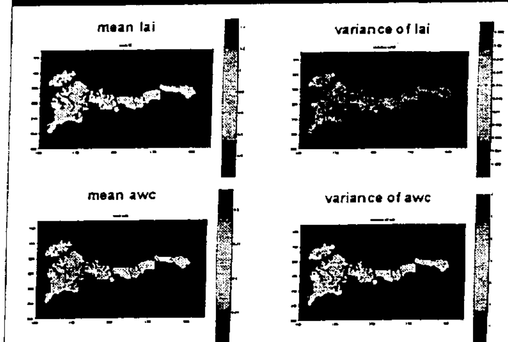
Results (cont):



Distribution of AWC from sampling, excluding the STATSGO prior

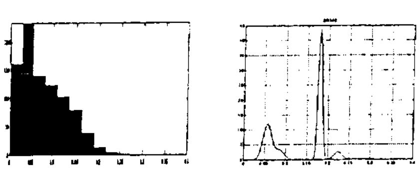


Results (cont):



Discussion:

- The awc distributions shown are for sampling p(ndvi | lai) p(lai | awc), ie not including the prior p(awc).
- Compare the distribution from sampling with the STATSGO prior:



- So the uncertainty is still on the level of the components.
- However: prior says that this component is only ~40% of the polygon, but the awc map seems to suggest that *all* of the pixels have approximately this distribution.

Further refinements:

- Improving p(ndvi | lai) and p(lai | awc) can be done by incorporating *other* types of information, eg
  - tree species,
    - age,
    - time since last clearcut 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

$$E(f(x)) = \int_x f(x)p(x)\partial x$$

- If we sample the domain of x uniformly, and compute p(x) for each x value, then the integral can be approximated by

$$E(f(x)) \approx \sum_i f(x(i))p(x(i))$$

- However this becomes inefficient if x is in more than a few dimensions (curse of dimensionality)
- Instead, if we sample from p(x), ie concentrate the samples in the high probabiltiy regions, then we can approximate the integral by

$$E(f(x)) \approx \sum_i f(x'(i))$$

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 Probabilties.*
- 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 $p_a = \min\left(1, \frac{prob(x' \mid Y)}{prob(x \mid Y)}\right)$
  - 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).