Experimental Design

Peter Rayner

March 16, 2015

Outline

- What are mathematicians good for anyway?
- Statistical background;
- First surprises;
- ED and making friends with the rich and powerless;
- Left-field interventions;
- What's happening now?
- A sober evaluation.

What are Mathematicians good for?

- "Let dead people do your work for you", only this time they weren't dead yet;
- Mathematicians see structural similarities in disparate fields;
- Ian, AGU 1992 was it?
- This little industry is built on judicious theft.

Statistical Background

$$p(x) \propto P(x|x^{\mathrm{b}}) \times p(y|H(x))$$

With x^{b} prior, y observations and H observation operator.

- Gaussian $p(x) \propto e^{-fracx-\mu 2\sigma^2}$;
- If H linear ($H(x) = \mathbf{H}x$) and rhs Gaussian then $p(x)$ also Gaussian (exercise for reader).

More statistical background

- Prior estimate x^{b} with uncertainty covariance $\mathbf{P}$;
- Data y with uncertainty covariance $\mathbf{R}$
- Linear observation operator $\mathbf{H}$;

$$x^{\mathrm{a}} = x\mathrm{b} + \mathbf{P} \times \mathbf{H}^{\mathrm{T}} \times \left[\mathbf{R} + \mathbf{HPH}^{\mathrm{T}}\right]^{-1} (y - \mathbf{H}x^{\mathrm{b}})$$

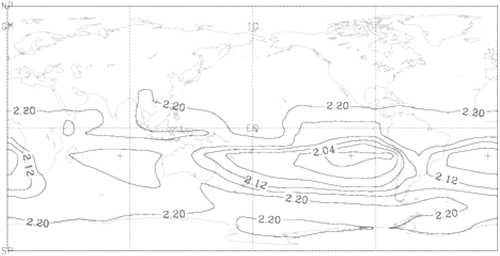
$$\mathbf{A}^{-1} = \mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$$

And the punch-line

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$$

- If we consider inverse covariance as information then posterior information = prior information + that projected back from data;
- Posterior covariances depend only on other covariances and observation operator, not values;
- Thus *we can calculate the value of measurements in reducing uncertainty without making them provided we know their characteristics.*;
- Can minimize some uncertainty metric wrt choice of observations.

First Network Design Map



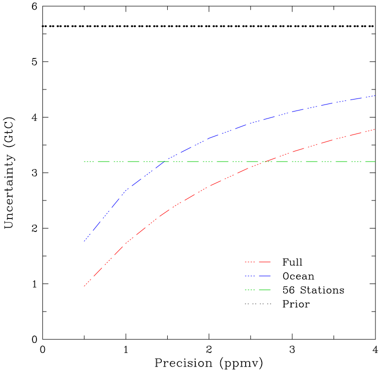
Standard deviation (in Gt!CY^{-1}) of global ocean uptake as a function of the position of one additional station to the current network. The station measures the same species with the same uncertainties as the existing station at Cape Grim, Tasmania. The contour interval is 0.02. The map also shows the positions of one and three additional stations calculated (with simulated annealing) to minimize the variance of net ocean uptake. Rayner et al., Tellus, 1996.

]

Most good ideas start life as a bug in the code

- Satellite data has *some* promise in carbon cycle inversions
- There is a limit to the usefulness of monthly mean data for determining even monthly mean sources
- Atmospheric transport imposes a relationship between optimal time and space scales and future sampling strategies should bear this in mind.

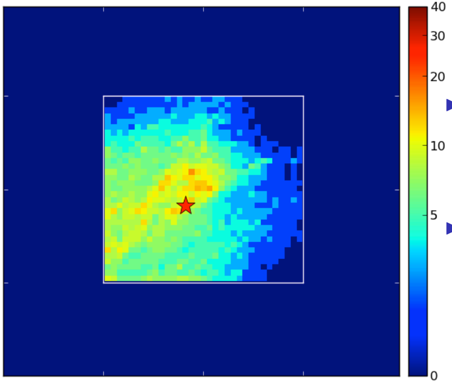
Impact of Global Satellite Measurements



Average uncertainty (GtCy^{-1}) against the precision (ppmv) of column-integrated data. The dotted horizontal line shows the prior uncertainty while the dash-dot horizontal line shows the case for the surface network.

- Rayner and O'Brien, GRL, 2001
- Surface network of 56 sites
- Monthly mean column-integrated data in every gridbox

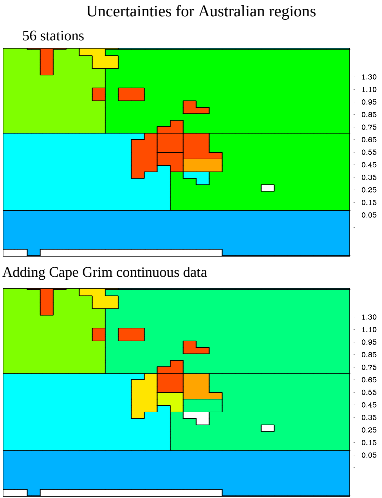
Regional Satellite Measurements



Uncertainty reduction ($100 * (1 - \sigma_{\mathrm{post}}/\sigma_{\mathrm{prior}})$ for emissions in idealized city. Case with CO and CO_2 measurements. Rayner et al., AMT (2014).

- Like prior since large uncertainties get larger reductions;
- Urban reductions stronger in northeast (follows from plume direction);
- Power-plant improved about 35%;
- No improvement on nighttime emissions.

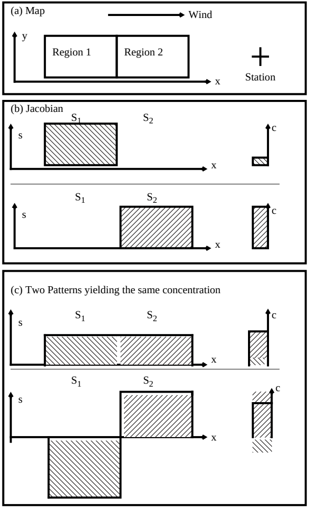
First continuous Data Experiment



- Continuous data appears to be useful.
- Most useful at smaller scales.
- Model rather than data quality may be limiting factor.

The Aggregation Problem

Patterns Determine Total Source



(a) An idealized situation of two regions upwind of an observational site. Since region 1 is farther upwind than region 2, for equal emissions S_1 and S_2, the source S_2 in region 2 has a higher impact on the concentration at the station. (b) Jacobian that quantifies this impact. (c) Two choices of emission patterns that yield the same concentration at the station.

Aggregation Consequences

- Trampert and Snieder, Science, 1996, Kaminski et al., JGR, 2001;
- Low resolution inversions may be biased by inhomogeneous sampling;
- Countered either by adding extra uncertainty or increasing resolution;
- Bocquet 2005 shows problem unstable as resolution $\rightarrow \infty$;
- Problematic term same as T&S uncertainty.
- Introducing prior correlations for smoothness reintroduces T&S truncation, problem unresolved.

Generalizations

- Posterior covariance embodies stats of posterior estimate with perturbed priors and data;
- Can capture these with Monte Carlo techniques if $\mathbf{P}$, $\mathbf{R}$ or $\mathbf{H}$ too big (e.g. Chevallier et al., JGR, 2007)
- This can also capture non-Gaussian or mistuned statistics (e.g. Chevallier et al., GRL, 2010, Burrows et al., ACP, 2013);
- Can linearize complex model with automatic differentiation then apply (e.g. Koffi et al., GBC, 2012).

Critiques and Responses

- The chosen network depends strongly on the uncertainty metric you minimize. *You'd hope the answer depends on the question*
- The prior uncertainty influences the network. *It's reasonable that where you measure depends on what you already know.*
- Data and model uncertainties are hard to characterize before you make measurements. *Err, Umm, next slide please.*

Characterizing Data Uncertainties

- Can model uncertainty, e.g. $\propto$ local flux
- Can treat small-scale variation as a model tracer (Galmarini et al., 2008);
- Can treat multiple models as an extra component of uncertainty to minimize.

How Useful has it Been?

- A great marketing tool, you can't propose a satellite these days without it;
- Has taught us a lot about general properties of networks e.g. coupling of land and ocean;
- Specific surface networks are so dominated by logistics that optimality usually gets pushed aside.