Statistical Modeling of Large Space-Time Datasets

Michael Stein

University of Chicago

Paris, June 2011

Funders & Collaborators

- US DOE and US NSF
- Mihai Anitescu, Stefano Castruccio, Jie Chen, Marcin Hitczenko, Feifei Liu, David McInerney, Liz Moyer

Topics

- Data
- Goals
- Models & Diagnostics
- Computation

Data

I prefer to think about modeling *processes* rather than *data*.

Nevertheless, the nature of the data available can have a major impact on

- the questions we can address
- the models we might use
- appropriate model diagnostics
- computational methods

Important aspects of space-time data:

- nature of response(s), e.g., percentage (relative humidity), vector (wind)
- frequency/extent in space and time
- regularity in space and time
- relation of measurements to quantity of interest

Irish Wind Data

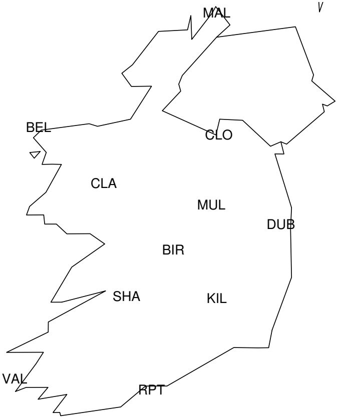
"Famous" (to those statisticians working on space-time data) dataset of daily wind speeds over 18 years at 12 sites in Ireland (Haslett and Raftery, 1989).

- No missing values!? Very convenient.
- One site doesn't seem to fit, so everyone drops it.
- If remove seasonal pattern, first differences (in time) of square root wind speed seem fairly close to stationary (in space-time) Gaussian process.

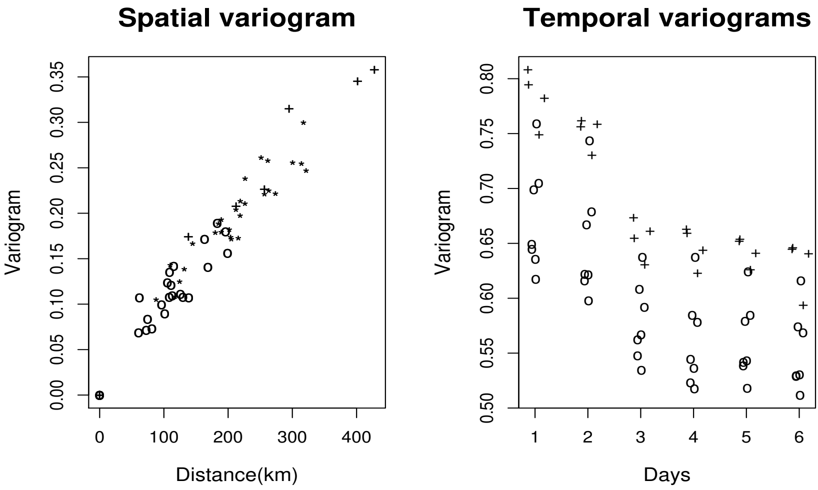
$72{,}314 = 11 \times 6{,}574$ observations no longer seems large and exact likelihood calculations should now be feasible.

Goal in Haslett and Raftery: Wind power prediction.

- Nonlinear (nonmonotonic!) function of wind speed. Are daily winds adequate?


MAL
BEL
CLO
CLA
MUL
DUB
BIR
SHA
KIL
VAL
RPT

Empirical variograms of first differences (+ = coastal, o = inland)



ARM SGP Data

Advanced Radiation Measurement (ARM) Southern Great Plains (SGP) site

- Established in 1992, "The SGP site is the largest and most extensive climate research field site in the world."
- Central Facility: many *in situ* and remote sensors as well as balloon-borne atmospheric profiling.
- Other facilities take less extensive measurements, including (as of now) 14 that measure surface meteorology every minute.
- Short-term field campaigns taking large amounts of specialized data (can produce over 2 gbytes/day).

The surface meteorological measurements are

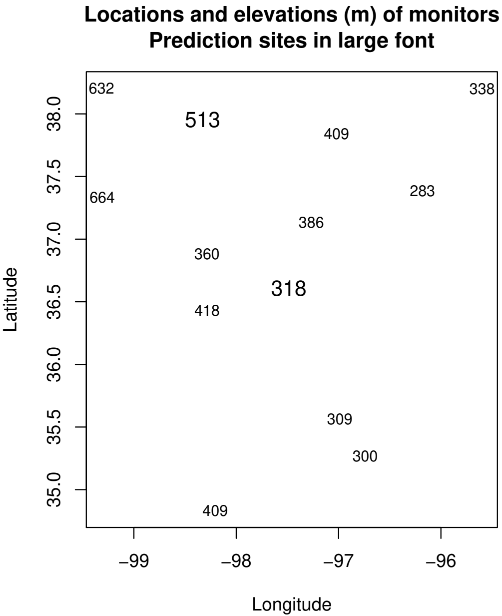
- Multivariate (temperature, humidity, pressure, winds)
- Regular and frequent in time (very low missing fraction)
- Sparse and irregular in space but at (largely) fixed sites

Over 3 million observations each year, so only moderately large by present standards.

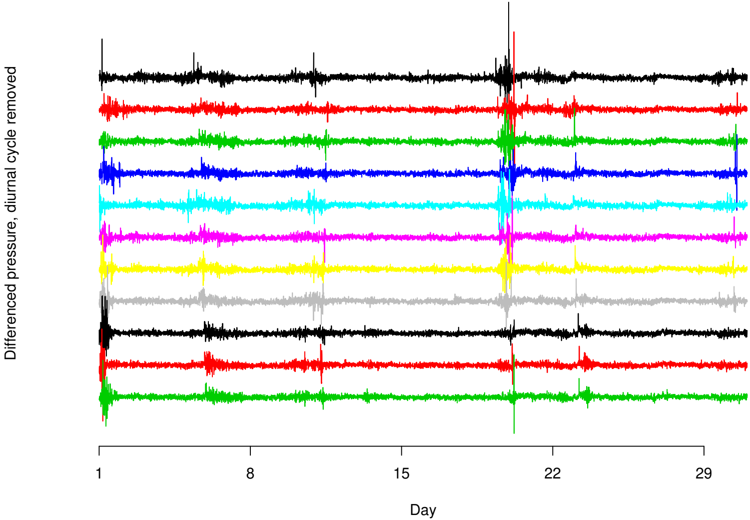
Goal: Highly resolved multivariate space-time conditional simulations of surface meteorology. Way too ambitious.

Easier but still challenging: five-minute averages of pressure for October, 2005 at 11 sites (predict pressure at 2 other sites).

- Clearly not stationary (in time) Gaussian process.


Locations and elevations (m) of monitors
Prediction sites in large font
632
338
513
409
664
283
386
360
318
418
309
300
409
Latitude
38.0
37.5
37.0
36.5
36.0
35.5
35.0
–99
–98
–97
–96
Longitude

Pressure differences at 11 sites, west to east



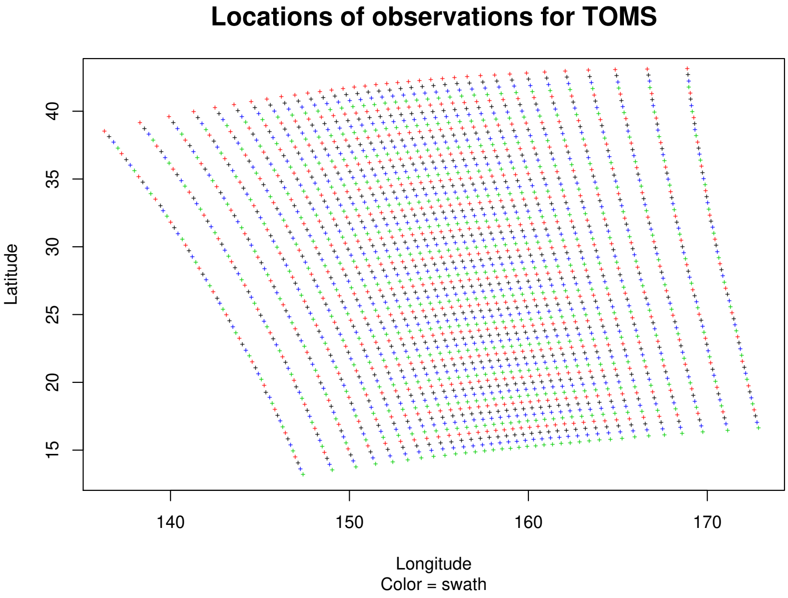
Level-2 TOMS

Total Ozone Mapping Spectrometer, based on a sun-synchronous polar-orbiting satellite.

- Daily measurements over 15 years with nearly global coverage (works on reflected light, so no data for polar nights).
    - About 180,000 observations (13.825 orbits/day $\times$ 378 swaths/orbit $\times$ 35 observations/swath) each day.
- Near equator, little overlap between scans on successive orbits, greater overlap away from equator.
- High resolution in space, but not in time.
- Data not on a grid in either space or time.

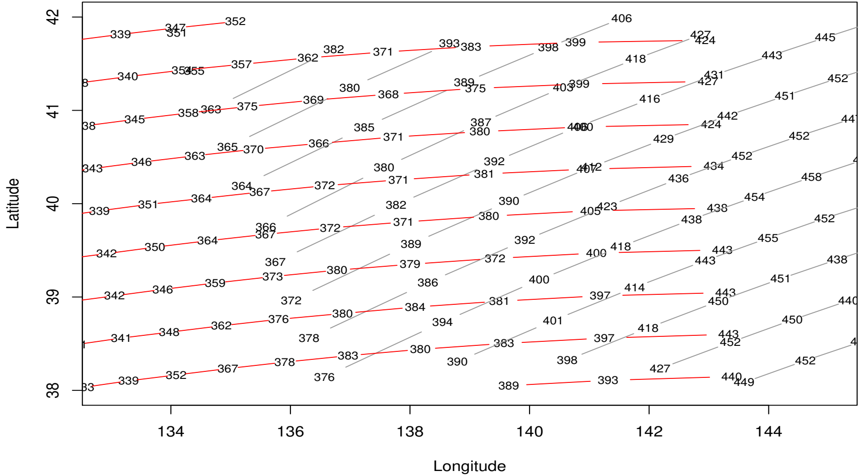
Consider 38°–42° N, May 1990.

Spatial variogram shows interesting feature: longitudinal asymmetry.

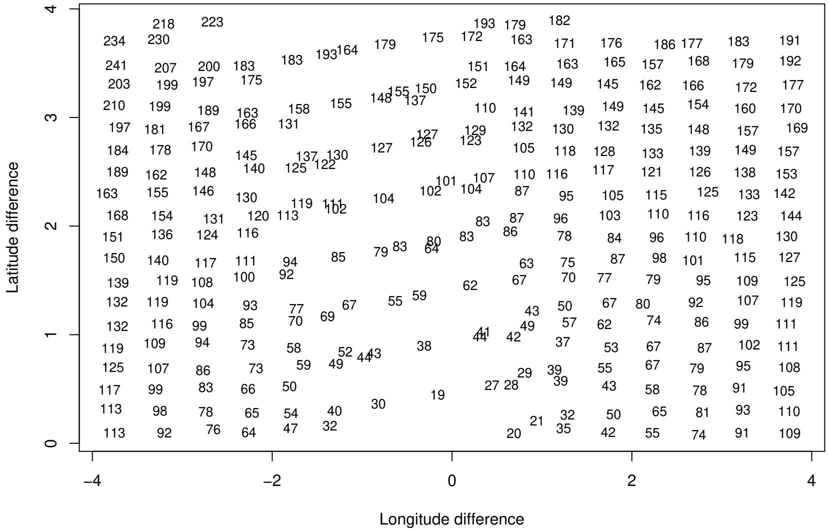


Locations of observations for TOMS
Latitude
15
20
25
30
35
40
140
150
160
170
Longitude
Color = swath

Some Level-2 TOMS on 5/1/1990



Empirical spatial variograms, May 1990, latitudes 38–42 N



Goal: Long-term trends at regional and seasonal scales.

- If data were complete, I would aggregate it to scales of interest before analyzing.
- Fair fraction missing; possible use for sophisticated statistical models?
  - Statistical models (or statisticians) not used to produce Level-3 TOMS.
- For atmospheric processes on a global scale, one often finds:
  - Process looks quite different at different latitudes.
  - Seasonal patterns depend on latitude.
  - Process behaves differently over land and water.
- Axially symmetric (Jones, 1963; Jun and Stein, 2007) a good place to start for global processes?

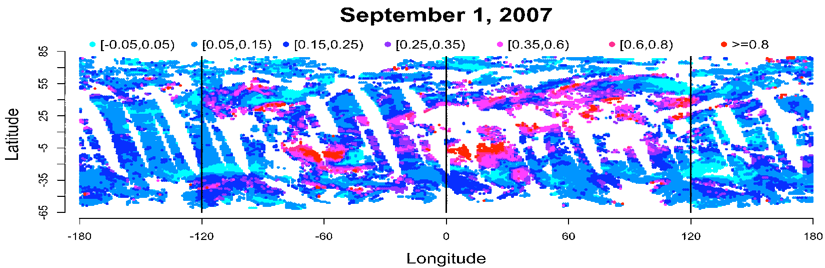
Aerosol optical depth

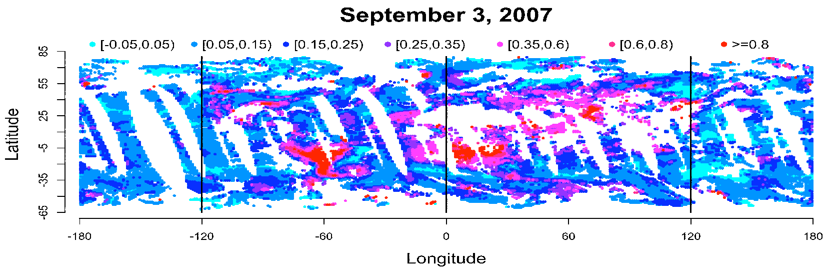
Marcin Hitczenko has analyzed 2 months of Level-2 aerosol optical depth measurements taken by MODIS (Moderate Resolution Imaging Spectroradiometer).

- 10 km spatial resolution but not quite global coverage on any one day.
- Generally no measurements over deserts or where there are clouds.
- About 600,000 observations a day.
- Data available on 1 km resolution as well.

Goals?

- Fill in gaps? But missingness not at random.
- Testbed for models and algorithms for statistical analysis of massive datasets.


September 1, 2007
[-0.05,0.05)
[0.05,0.15)
[0.15,0.25)
[0.25,0.35)
[0.35,0.6)
[0.6,0.8)
>=0.8
Latitude
85
55
25
-5
-35
-65
-180
-120
-60
0
60
120
180
Longitude


September 3, 2007
[-0.05,0.05)
[0.05,0.15)
[0.15,0.25)
[0.25,0.35)
[0.35,0.6)
[0.6,0.8)
>=0.8
Latitude
85
55
25
-5
-35
-65
-180
-120
-60
0
60
120
180
Longitude

Climate model output

RDCEP: A Center for Robust Decision Making on Climate and Energy Policy

- An NSF supported center for improving models used to forecast the impact of policies on future economic and climatic conditions.
- Focus on economic models, but climate affects economies.
- Model for world economy generates emissions scenario as function of policies.
  - Climate forecasts for broad range of emissions scenarios must be quickly computable to be usable within the economic models.
  - *Cannot* run any nontrivial climate model for every emissions scenario of interest.

Need fast "emulator" of climate model output (or better yet, of actual future climates).

For some moderate number of scenarios of CO_2 trajectories (not emissions), we can run CCSM3.0:

- NCAR climate model released in 2004, so sophisticated but not quite cutting edge.
  - Coupled atmosphere/ice/land/ocean model.
  - Needs initial conditions and greenhouse gas concentrations as inputs.
- Because of sensitivity to initial conditions, unlike many computer models, output is effectively stochastic.

Goal: Emulate CCSM3.0 output for any plausible future CO_2 trajectory.

- $C(x, t; \mathrm{CO}_2, \mathrm{IC})$ is the temperature and precipitation from CCSM3.0 at (x, t) for some CO_2 scenario and initial conditions IC.
- Have $C(x, t; \mathrm{CO}_2, \mathrm{IC})$ for some set of $(\mathrm{CO}_2, \mathrm{IC})$ scenarios.
- Since IC "unknown," forecast multivariate "distribution" of $C(x, t; \mathrm{CO}_2, \mathrm{IC})$ over IC given CO_2.

Initial plan: view the multivariate climate output as a Gaussian process depending on the annual CO_2 levels.

Using entire CO_2 trajectory as input is problematic:

- Uses CO_2 after time t to forecast climate at time t.
- To get prediction at time t, could restrict to using trajectories only up to time t for both "training" and "test" runs.
  - Requires using different set of inputs for every t.
  - Throws away relevant information from "training" runs.
  - Tracking time by calendar year is just wrong.
  - Ignores likely monotonic effect of past CO_2 on temperature.

Alternative: Think of input as back trajectory of CO_2.

Regression model for, say, temperature:

$$T(x,t;\mathrm{CO}_2,\mathrm{IC}) = \sum_{j=1}^{p} f_j(\mathrm{CO}_2(t),\ldots,\mathrm{CO}_2(t-T);\theta_j(x)) + e(x,t;\mathrm{CO}_2,\mathrm{IC}),$$

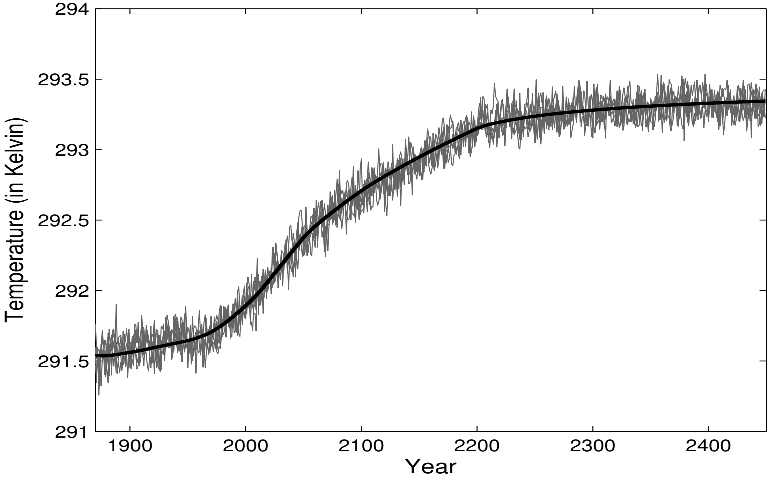
for e some space-time random field that is independent for different $(\mathrm{CO}_2,\mathrm{IC})$.

- Possible $f_j(\mathrm{CO}_2(t),\ldots,\mathrm{CO}_2(t-T);\theta_j(x))$:
  - $\theta(x)\log\{\mathrm{CO}_2(t)\}$
  - $\theta_1(x)\sum_{s=0}^{T} e^{-\theta_2(x)s}\log\{\mathrm{CO}_2(t-s)\}$
- Makes explicit use of time order.
- Takes account of processes on different time scales.

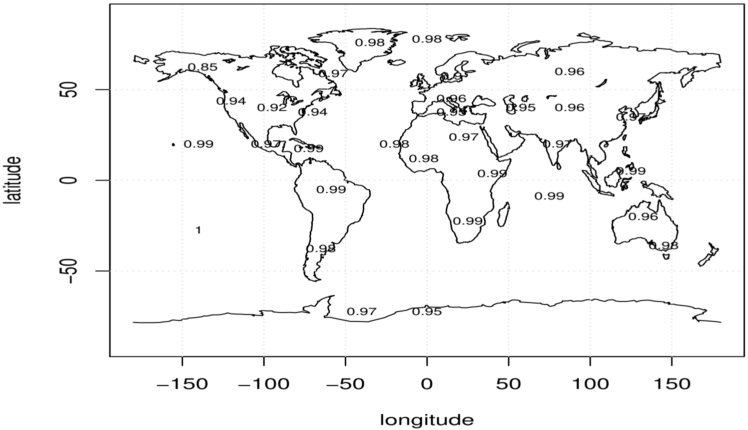
Time scales (e.g., $\theta_2(x)$) over land v. water quite different.

Example: use runs with quickly and slowly increasing CO_2 to predict for moderately increasing scenario.

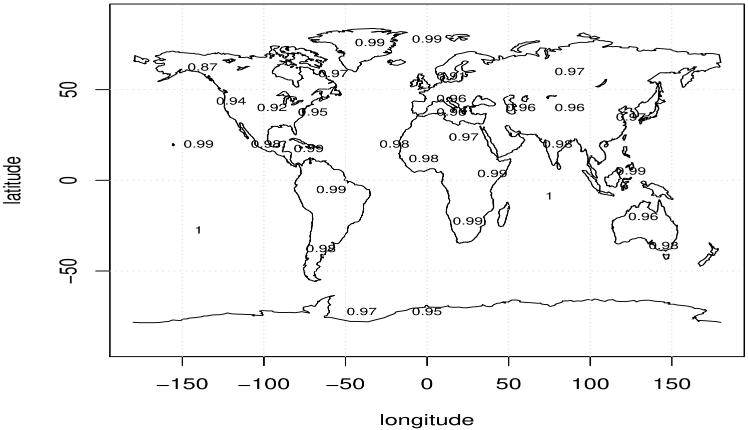
Fitted and five realized temperature series in South Pacific



R^2 values for average of five runs by region based on low and high runs



R^2 values for moving average fit



Models & diagnostics

Diagnostics require at least an implicit model to diagnose (e.g., stationary).

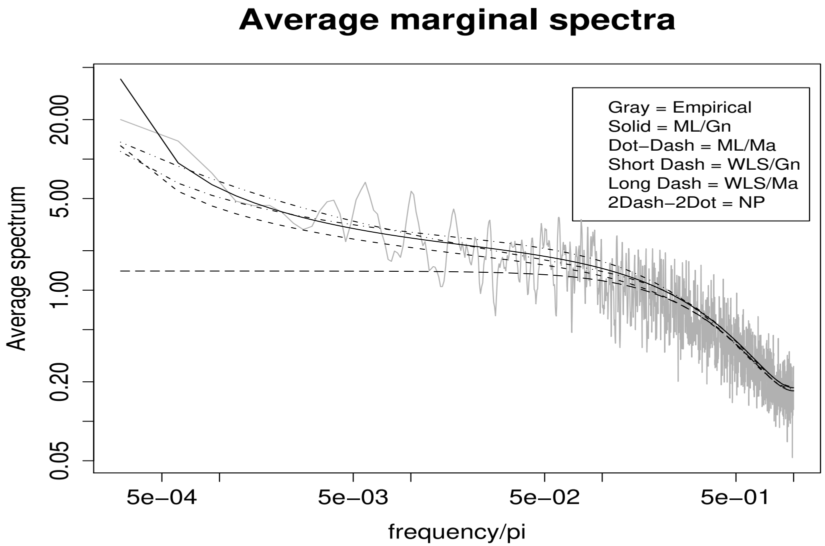
Points in a diagnostic plot should have known simple structure when proposed model is correct; e.g., independent and identically distributed.

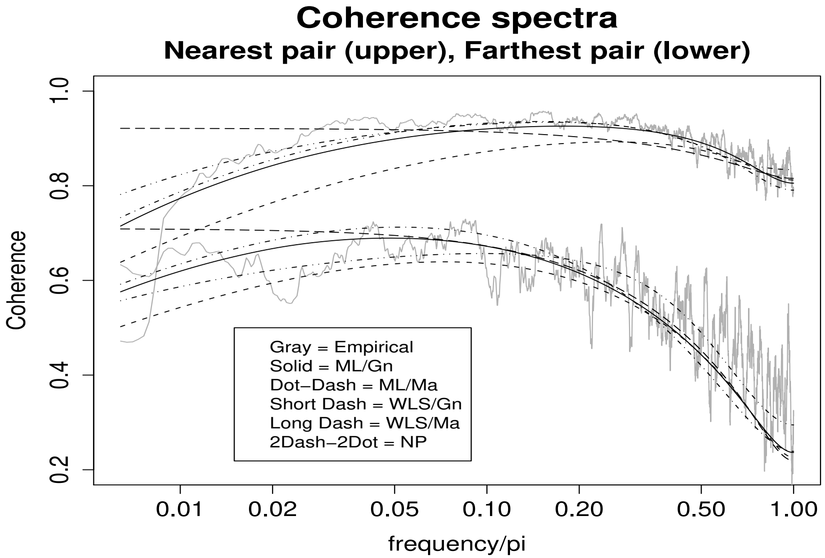
- If not identically distributed, then at least nearby points should be close to identically distributed.
- If not independent, then at least not too dependent.

The empirical variogram is a valuable diagnostic for spatial and spatial-temporal data, but

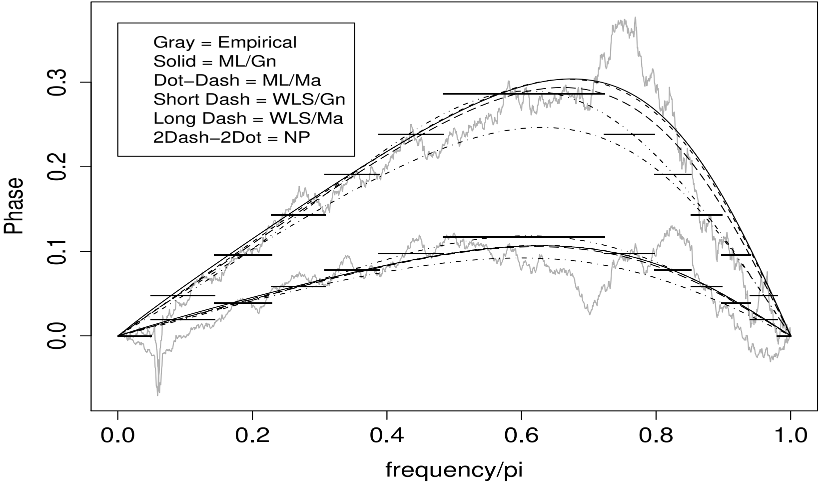
- values at similar lags can be highly correlated
- space-time setting opens new issues beyond purely spatial setting

For Irish wind data, data structure suggests use of spectral methods.


Average marginal spectra
Gray = Empirical
Solid = ML/Gn
Dot−Dash = ML/Ma
Short Dash = WLS/Gn
Long Dash = WLS/Ma
2Dash−2Dot = NP
Average spectrum
20.00
5.00
1.00
0.20
0.05
5e−04
5e−03
5e−02
5e−01
frequency/pi


Coherence spectra
Nearest pair (upper), Farthest pair (lower)
Coherence
1.0
0.8
0.6
0.4
0.2
0.01
0.02
0.05
0.10
0.20
0.50
1.00
frequency/pi
Gray = Empirical
Solid = ML/Gn
Dot−Dash = ML/Ma
Short Dash = WLS/Gn
Long Dash = WLS/Ma
2Dash−2Dot = NP

Phase spectra with Dublin
Claremorris (upper), Mullingar (lower)



Diagnostics in space-time domain

Let

- $\mathbf{Z}_t = n$-vector of observations at time t
- $\mathbf{b} = n$-vector of coefficients

For $k = 1, 2, \ldots$ define

$$\bar{\mathbf{Z}}_{j,k} = \frac{1}{k} \sum_{\ell=j}^{j+k-1} \mathbf{Z}_\ell$$

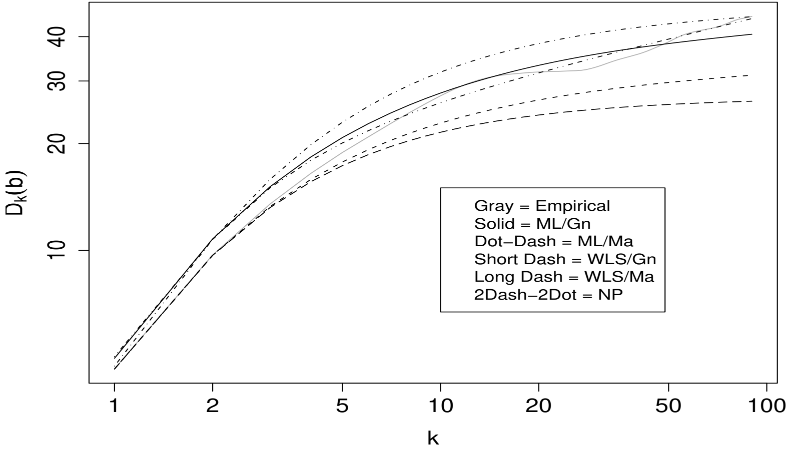
and

$$D_k(\mathbf{b}) = \frac{1}{T - 2k + 1} \sum_{j=1}^{T-2k+1} \left\{ \mathbf{b}'(\bar{\mathbf{Z}}_{j,k} - \bar{\mathbf{Z}}_{j+k,k}) \right\}^2$$

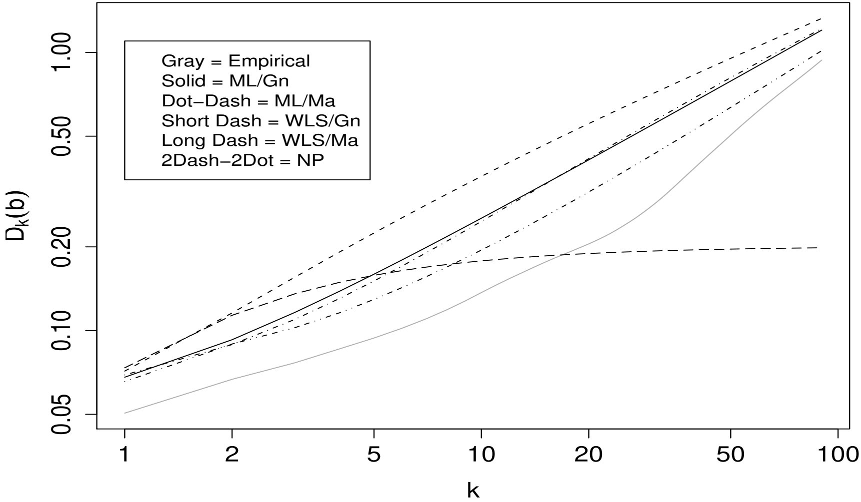
for $\mathbf{b} = 11^{-1/2}\mathbf{1}$ (spatial average) and

$$\begin{aligned} \mathbf{b} &= \text{contrast eliminating linear polynomials for 4 nearby sites} \\ &= 0.40 \text{ BIR } + 0.15 \text{ DUB } - 0.85 \text{ MUL } + 0.40 \text{ CLO} \end{aligned}$$

Spatial average



Contrast eliminating linear polynomials



Diagnostics for TOMS data

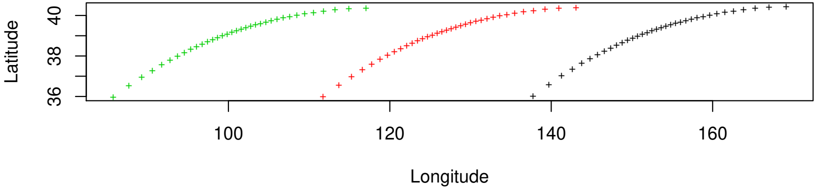
TOMS data close to axially symmetric: spatial variation nearly invariant to rotations about Earth's axis.

Consequence: For $\mathbf{H}$ a rotation matrix,

$$\mathrm{var}\left\{\sum_{j=1}^{n} \lambda_j Z(\mathbf{H}\mathbf{x}_j)\right\}$$

depends on $\mathbf{H}$, but not so much if $\mathbf{H}$ is a rotation about Earth's axis.

Example: For each of 82 orbits in March 1–6, 1990, consider the first swath with first observation above latitude 40°N.



Yields 82 sets of 35 observations whose locations from one orbit to the next are nearly rotations about Earth's axis of each other.
If $\mathbf{x}_1, \ldots, \mathbf{x}_{35}$ are locations of swath in first orbit, have 82 not quite identically distributed and (not too?) dependent estimates of

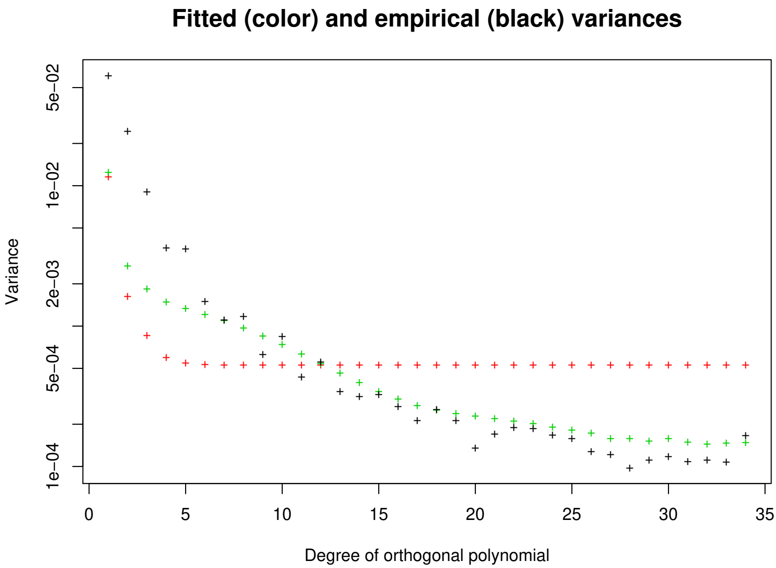
$$\mathrm{var}\left\{\sum_{j=1}^{35} \lambda_j Z(\mathbf{x}_j)\right\}.$$

Choosing λ_j's: Orthogonal polynomials of degrees 1–34 (treat observations as if evenly spaced on a line).

- Like spectral analysis.
- If truth is white noise, variances of all 34 contrasts the same.

Observed versus fitted (mle) values for these 34 variances.

- Red model uses nugget + reduced rank covariance function.
- Green model adds a compactly supported component.


Fitted (color) and empirical (black) variances
Variance
5e−02
1e−02
2e−03
5e−04
1e−04
0
5
10
15
20
25
30
35
Degree of orthogonal polynomial

Spectral in time modeling approach

Parametric v. nonparametric approaches to modeling.

- Want greater flexibility for those aspects of model about which have most information.
- For Irish wind and ARM data, have lots of replication in time.

Example (Stein, 2005):

$$K(\mathbf{x}, t) = \int_{-\pi}^{\pi} S(\omega) C\left(|\mathbf{x}|/\delta(\omega)\right) e^{i\theta(\omega)\mathbf{u}'\mathbf{x}+it\omega} d\omega$$

is positive definite and real on $\mathbb{R}^d \times \mathbb{Z}$ if

- S is even and integrable
- δ is even and θ is odd
- C is a valid isotropic spatial correlation function

Interpretation of

$$\int_{-\pi}^{\pi} S(\omega) C\left(|\mathbf{x}|/\delta(\omega)\right) e^{i\theta(\omega)\mathbf{u}'\mathbf{x}+it\omega} d\omega.$$

- S is spectral density in time at any site
- δ controls coherence
- θ controls phase

Irish wind data:

- "NP" model in diagnostic plots.
  - Fit by approximate likelihood in spectral domain.
- Other models are elaborate parametric models.
  - Fits via (approximate) mle or wls fits to various empirical variograms.

Computation

Exact computations (kriging, Gaussian likelihoods) for large, irregularly sited datasets generally requires $O(n^3)$ computation and $O(n^2)$ memory.

Options for large n:

- Use model that reduces computation and/or storage.
- Use approximate methods.
- Both.

Now working on project with "petascale" ($n \approx 10^{15}$) in title.

- Even for terascale ($n \approx 10^{12}$) data, probably need single-pass methods if want to fit global model.

Models that reduce computation

Compactly supported covariance functions

- Spherical, models in Gaspari and Cohn (1999).
- Produces sparse matrices, which reduces storage and computations.
  - Sparseness easily exploitable for solving linear systems.
  - Not so easy to exploit for log determinants (location of zeroes matters).
- Can cause problems:
  - Lack of screening effect.
  - Lack of differentiability of likelihood with respect to range.
- Despite their benefits, I don't think they are the best approach.

Reduced rank covariance functions (Cressie and collaborators):

$$\text{cov}\{Z(x), Z(y)\} = \text{nugget} + \sum_{j=1}^{m} a_j b_j(x) b_j(y)$$

for a_j's nonnegative.
If m is much smaller than sample size, great (and easy) reduction in storage and computation (including log determinant).

- Problems modeling local behavior, especially when nugget is modest compared to variation between neighboring observations (TOMS, MODIS).
  - Likelihood estimates may give terrible match for empirical variogram.

Stein (2007) added a covariance function with quite narrow support to address this problem for TOMS data ($+$'s).

- Helped quite a bit, but still some clear misfit.

Markov models (MRFs, Kalman filter for space-time setting).

Approximate computation

For massive, irregularly sited datasets, approximate computation is unavoidable (although see Katzfuss and Cressie, 2011).

- Just fit models locally.
- Spectral methods (Whittle likelihood).
  - Best for gridded data from stationary processes.
- Various forms of composite likelihood:
  - Write joint density as product using successive conditioning; condition on only part of "past" (Vecchia 1988; Stein, Chi and Welty 2004).
  - Combine local and sparse subsets of data (Carragea and Smith).
- Covariance tapering (Furrer, Genton and Nychka, 2006; Kaufman, Schervish and Nychka, 2008; Loh and Wang, 2009).

Covariance tapering straddles change the model/change the computation divide:

- Multiply (elementwise) covariance matrix of interest by sparse covariance matrix.
- If matrices $K, T \geq 0$ then $K \circ T = (k_{ij} t_{ij}) \geq 0$.
- For a dense matrix K, try to find sparse T so that $K \circ T$ gives similar inferences as K.
  - Example: K and T have spectral densities f and τ with τ/f sufficiently small at high frequencies.
- Either act as if $K \circ T$ is truth (change the model) or use estimating equations approach (change the computation).

Interesting application of theory (equivalence of Gaussian measures) to computational problem.

For massive datasets with strong correlations, need something more?

Covariance tapering can be applied to *any* positive definite matrix.

- So first *filter* the data to reduce the correlations and *then* taper?
  - Not so clear how to do this with irregularly sited observations.

Convergence of iterative methods for solving linear equations related to condition number $\kappa(K)$ of covariance matrix K.

Result from Stein, Chen and Anitescu (unpublished):
Z on real line with spectral density f satisfying

$$f(\omega)\omega^4 \text{ bounded away from 0 and } \infty \text{ as } \omega \to \infty.$$

Let L be filter matrix for normalized second differences.

There exists $C_f < \infty$ such that, for any set of observations of Z in $[0, 1]$,

$$\kappa(LKL^T) \leq C_f.$$

Maximum likelihood estimates

Optimization methods such as conjugate gradient require derivatives.

- If having numerical problems, compute first derivatives analytically.

Hessian useful to scale components of parameter vector.

- In high dimensions, this scaling sometimes essential.
- Even crude approximations to Hessian may be adequate.

Hitczenko developed methods to do this with processed MODIS data to fit axially symmetric models with many parameters.

- Even so, required parallel computation to be feasible.
- Despite huge effort, still had poor fit to local variation.

Maybe we don't need to compute likelihoods to find mle?

- Solve score equations instead?

For covariance matrix $K(\theta)$, requires

- Quadratic forms (relatively easy)
- For each component of θ,

$$\mathrm{tr}\left\{K(\theta)^{-1}\frac{\partial}{\partial\theta_i}K(\theta)\right\} \approx \frac{1}{N}\sum_{j=1}^{N} u_j^T K(\theta)^{-1}\frac{\partial}{\partial\theta_i}K(\theta)u_j,$$

where u_j's have components ± 1, each with probability $\frac{1}{2}$.

- If pick u_j's well, can get away with N quite small (at least much smaller than sample size)?

Uniqueness of solution?

One-pass methods

Look at data block by block and summarize the information about $K(\theta)$ from that block so that don't have to go back to again.

Simple example:

- Divide data into B blocks.
- Within each block, find mle of θ and observed information matrix.
  - Gives a quadratic approximation to loglikelihood within each block.
- Also save "corner" observations from each block.
- Add within block approximate loglikelihoods to loglikelihood of all corner observations.

When might this procedure do asymptotically as well as full likelihood?

Other critical topics

- Nonstationary models. Axially symmetric model an example?
- Models for measurement processes (remote sensing).
- Space is three-dimenional, not two. Atmospheric processes change character with altitude.
- Simultaneous modeling of physically linked quantities like wind and pressure or temperature and relative humidity.
- Getting more (but not too much?) science into statistical models (Cressie and Wikle, 2011).

Some observations:

- We live in a world indexed by space and time.
- The biggest scientific and policy questions increasingly involve issues of difficult to characterize uncertainty and variability.
- Statistical methods for space-time data are in their infancy.

Conclusion: We have a lot of work to do!