Spatio–temporal methods in environmental epidemiology: Lecture 24

Lecture 4

SPECIAL TOPICS:

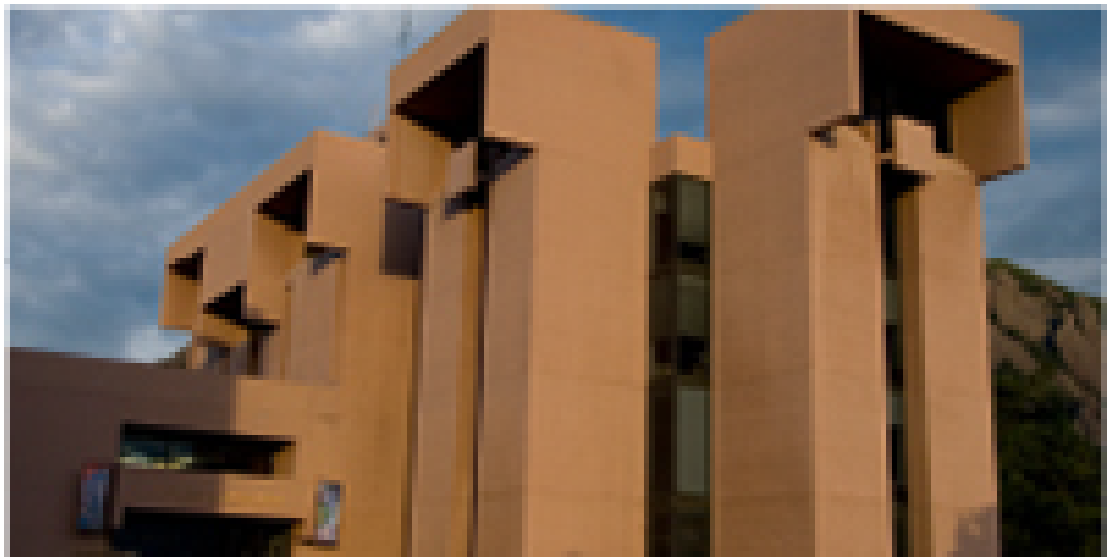
Big data analysis (BDA) & High dimensional data analysis (HDDA)

BDA

Big data analysis (or “analytics”) is the process of examining large amounts of data of a variety of types (big data) to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in business benefits, such as more effective marketing and increased revenue. Can also be used to seen spatio–temporal trends in climate characteristics.

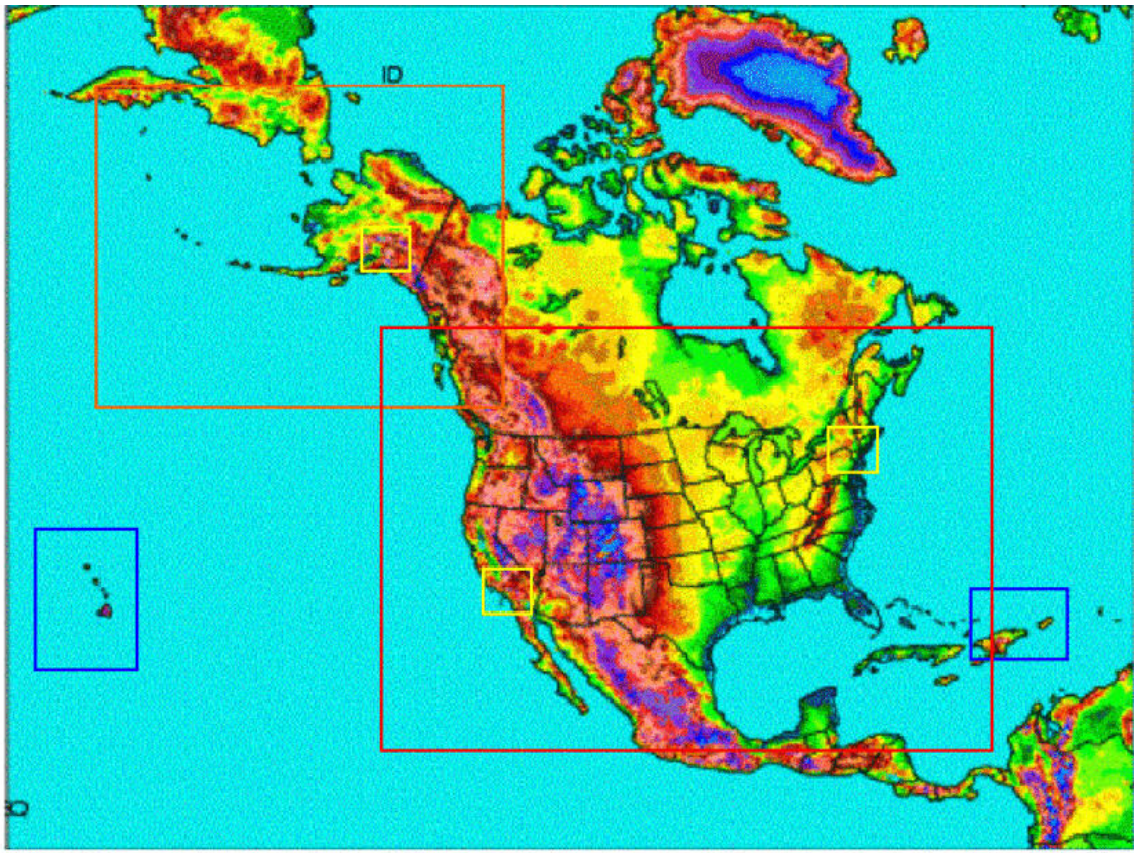
Involves storage, retrieval, high powered processors and fast computational approaches. A must know for modern statistical scientists.

Example: The US National Center for Atmospheric Research (NCAR) in Boulder, Colorado receives a terabyte of data per day of climate data from around the world.



NCAR Data storage racks - water cooled - slaves managed by 5 master servers with redundancy built-in

These vast stores have to be processed quickly for climate analyses



HDDA

High dimensional data analysis refers to data and or the processes that generate them, where very high dimensional vectors are involved. The global network of temperature monitors would yield data vectors like that. The "p vs n problem" arises where p may be the dimension of the data vector and n is the number observed with $n << p$.

1. Global temperature monitors.
2. Hi - dimensional data vectors (Z) : Data are functions recorded on monitors over a continuous spectrum that could include frequencies (nuclear magnetic resonance analyses) for a few specimens or subjects.

Responses may be highly correlated & need arises for data compression like extracting the principal components of variation from specimen–to–specimen.

Latent responses (Y) or parameter vectors (β) :

$$Z^{n\times p} = K^{n\times q} Y^{q^{\times} p} + \epsilon^{n^{\times} p}, \text{ or } Z = X\beta + \epsilon$$

Partially observed processes (Y):

$$Y^{n\times p} = (Y^{n\times u}, Z^{n\times g})$$

Dealing with high dimensions

- High compression methods
  - principal components. What if the data matrix is an array and not a vectors?
  - empirical orthogonal functions. Like PCA but for spatial patterns
- Use Laplace-like approximations to avoid MCMC and Winbugs, e.g. Enviro.stat and Integrated Nested Laplace Approximation [1] (INLA).
- Etc.

[1] Cameletti et al. [2011],Lindgren et al. [2011],Rue et al. [2009]

First: conventional Bayesian approach [2]

The process is monitored at sites $\{s_1, \ldots, s_g\}$

Data model: $Z_t(s_i) = x_t^{\text{Covariates}}(s_i)\beta + Y_t^{PM_{10}\ \text{process}}(s_i) + \epsilon_t(s_i),\ i = 1, \ldots, g$

Process model: $Y_t(s_i) = aY_{t-1}(s_i) + \omega_t(s_i),\ t = 1, \ldots, T, i = 1, \ldots, g, g+1, \ldots, g+u = p$

Parameter model: $Cov(\omega_t(s_i), \omega_{t'}(s_i)) = \sigma_\omega^2 C(h)\ \mathrm{I}\{t = t'\}$ where C denotes the Matern covariance model

[2] Cameletti et al. [2011]

NOTES:

(1.) The process model covers all *p* prospective sites. This is needed for spatial prediction.

(2.) The $\{\omega_t\}$ do not carry any of the temporal correlation - that is left to the AR(1) process model. Instead their job is to carry the spatial correlation.

(3.) The coefficient *a* does not depend on time - a potential limitation for say hourly data where the vector approach of enviro.stat would work better and help avoid potential spatial correlation leakage.

(4.) The $\{\epsilon_t\}$ represent classical measurement errors and are normally distributed.

NOTES:

(5.) **The matern covariance for a stationary Gaussian process:** for locations u, v the Matern covariance between responses is

$$C(h) = \sigma^2(\kappa h)^\nu K_\nu(\kappa h)/2^{\nu-1}\Gamma(\nu).$$

for a stationary field with $h = || u - v ||$, where K_ν is modified Bessel function of 2nd kind; σ^2 is overall variance; (ν, κ) are smoothness and range parameters.

NOTES:

(6.) **Vector forms:**

$$
\begin{aligned}
Y_t &= (Y^{(g)}, Y^{(u)}) \\
Y^{(g)} &= (Y_t(s_1), \ldots, Y_t(s_g)), \text{ etc.} \\
Z_t &= x_t\beta + Y_t^{(g)} + \epsilon_t, \ \epsilon_t \sim N(0, \sigma_\epsilon^2 I_g) \\
Y_t &= aY_{t-1} + \omega_t, \ \omega_t \sim N(0, \Sigma = \sigma_\omega^2 \tilde{\Sigma}) \\
Y_1 &\sim N(0, \Sigma/(1-a^2))
\end{aligned}
$$

Posterior distributions For simplicity drop superscripts "(g)". Then $\theta = (\beta, \sigma_\epsilon^2, \sigma_\omega^2, a, \kappa)$ has joint posterior density

$$\begin{aligned}
\pi(\theta, y_t \mid z) &\propto \pi(z \mid y, \theta)\pi(y \mid \theta)\pi(\theta) \\
&\propto \left(\Pi_{t=1}^{T} \pi(z_t \mid y_t, \theta)\right) \times \\
&\quad \left(\pi(y_1 \mid \theta)\Pi_{t=2}^{T} \pi(y_t \mid y_{t-1}, \theta)\right) \pi(\theta)
\end{aligned}$$

More explicitly

$$
\begin{aligned}
\pi(\theta, y \mid z) \quad \propto \quad & (\sigma_\epsilon^2)^{-\frac{gT}{2}} \exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^{T}(z_t - x_t\beta - y_t)'(z_t - x_t\beta - y_t)\right) \\
& \times \left(\frac{\sigma_\epsilon^2}{1-a^2}\right)^{-\frac{g}{2}} \mid \tilde{\Sigma} \mid^{-\frac{1}{2}} \exp\left(-\frac{(1-a^2)}{2\sigma_\epsilon^2} y_1' \tilde{\Sigma}^{-1} y_1\right) \\
& \times (\sigma_\epsilon^2)^{\frac{-g(T-1)}{2}} \mid \tilde{\Sigma} \mid^{-\frac{(T-1)}{2}} \\
& \times \exp\left(-\frac{1}{2\sigma_\omega^2}\sum_{t=2}^{T}(y_t - a y_{t-1})'\tilde{\Sigma}^{-1}(y_t - a y_{t-1})\right) \\
& \times \Pi_{i=1}^{dim(\theta)} \pi(\theta_i)
\end{aligned}
$$

Clouds on the horizon

Lack of explicit form Makes model properties hard to assess since all you have are immense MCMC samples and their empirical assessments. In particular, degree of correlation leakage.

Computational issues MCMC sampling will not be practical when p exceeds say 500. (**Exercise: Invert a 700 $\times$ 700 positive definite matrix in R!**). Suggests use of approximations to the posterior distribution, such as that of enviro.stat (empirical Bayes) or a Laplace approximation (Integrated nested Laplace approximation – **INLA**

Laplace (1749 –1827): Brilliant young scholar - admitted to the French Academy at age 24! Rediscovered Bayes amongst other things.

$$F(s)=\int_{0_-}^{\infty}\{\cos(\omega t-\varphi)\}\exp(-st)\,dt \qquad \Rightarrow \begin{cases}\cos(\omega t) & \text{when } \varphi=0\\ \sin(\omega t) & \text{when } \varphi=\frac{\pi}{2}\end{cases}$$

$$=\int_{0_-}^{\infty}\frac{1}{2}\left\{\exp\left[(j\omega-s)t-j\varphi\right]+\exp\left[-(j\omega+s)t+j\varphi\right]\right\}dt$$

$$=\frac{1}{2}\left\{\frac{1}{(j\omega-s)}\exp\left[(j\omega-s)t-j\varphi\right]_{0_-}^{\infty}+\frac{1}{-(j\omega+s)}\exp\left[-(j\omega+s)t+j\varphi\right]_{0_-}^{\infty}\right\}$$

$$=\frac{1}{2}\left\{\frac{-1}{(j\omega-s)}\exp(-j\varphi)+\frac{1}{(j\omega+s)}\exp(+j\varphi)\right\}$$

$$=\begin{cases}\frac{1}{2}\left[\frac{-1}{(j\omega-s)}+\frac{1}{(j\omega+s)}\right]=\frac{s}{s^2+\omega^2} & \text{for } \varphi=0\\ \frac{j}{2}\left[\frac{1}{(j\omega-s)}+\frac{1}{(j\omega+s)}\right]=\frac{\omega}{s^2+\omega^2} & \text{for } \varphi=\frac{\pi}{2}\end{cases}$$

Laplace approximation: General case where θ has dimension d:

$$\begin{aligned} f(y) &= \int \pi(y \mid \theta)\pi(\theta)d\theta \\ &= \int \exp\{-nh(\theta)\}d\theta \\ &= \pi(y \mid \hat{\theta})\pi(\hat{\theta})(2\pi)^{(d/2)} \mid \Sigma \mid^{1/2} n^{-1/2} \end{aligned}$$

where

$$\begin{aligned} h(\theta) &= -\frac{1}{n}\log \pi(y \mid \theta) - \frac{1}{n}\log \pi(\theta) \\ \Sigma &= (D^2 h(\hat{\theta}))^{-1}, \text{ Inverse Hessian matrix} \end{aligned}$$

INLA approach

- Originally developed for lattice processes with non-Gaussian measurement models, e.g. Poisson [3]:

$$\pi(y_i \mid z) = \int \pi(y_i \mid \theta, z)\pi(\theta \mid z)d\theta$$

$$\pi(\theta_j \mid z) = \int \pi(\theta_j \mid z)d\theta_{-j}$$

is replaced by

$$\tilde{\pi}(y_i \mid z) = \int \tilde{\pi}(y_i \mid \theta, z)\tilde{\pi}(\theta \mid z)d\theta$$

$$\tilde{\pi}(\theta_j \mid z) = \int \tilde{\pi}(\theta_j \mid z)d\theta_{-j}$$

where s means Laplace approximations nested within the integrals.

[3] Rue et al. [2009]

The approximations use such things as:

$$\tilde{\pi}(\theta_j \mid z) \propto \left. \frac{\pi(y, \theta, z)}{\tilde{\pi}(y \mid \theta, z)} \right|_{y=y^*(\theta)}$$

Extending INLA to point referenced data[4]

Step 1. Converts point referenced data to Markov random field (MRF)

Step 2. Exploit MRF neighbourhood structure to get sparse covariance matrices for Gaussian MRFs (GMRFs) and easy inversion

Step 3. Use INLA to avoid fully Bayes and MCMC methods

[4] Lindgren et al. [2011]

INLA building blocks: Sparse covariance matrices

Starting point: **multivariate normal distribution (MVN)**. General theory:

$$U^d \sim N_d(\mu, \Sigma)$$

means

$$E[U \mid \mu, \Sigma] = \mu; Cov(U_i, U_j \mid \mu, \Sigma) = \Sigma_{ij}$$

Let $Q = \Sigma^{-1}$ be the "precision matrix" of U. Then (**Exercise**)

$$Cov(U_i, U_j \mid \mu, \Sigma, U_{-\{ij\}} = u_{-\{ij\}}) = 0 \Leftrightarrow Q_{ij} = 0.$$

This means

$$Q_{ij} \neq 0, j \in \{i, N(i)\}$$

where $N(i)$ is the neighbourhood of i in any GMRF U

Implications: one can map out the neighbourhood structure in a GMRF U through Q. A big discovery in the early work on lattice–based processes. If for example $d = 500,000$, Σ could not be inverted numerically but neighbourhood structure could be used to build a Q, most of whose elements are 0 (a sparse matrix). A GMRF would then be practical.

INLA building blocks: Matern & spatial fields

Suppose Gaussian random field $U(s)$ on R^d has Matern covariance:

$$C(h) = \sigma^2(\kappa h)^\nu K_\nu(\kappa h)/2^{\nu-1}\Gamma(\nu).$$

Then U is solution of [5] a stochastic partial differential equation (SPDE):

$$(\kappa^2 - \Delta)^{\alpha/2} U(\mathbf{u}) = \mathcal{W}(\mathbf{u}),\ \mathbf{u} \in \mathcal{R}^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0$$

where $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudo – differential operator, Δ is the Laplacian, $\mathcal{W}$ is spatial white noise with unit variance.

[5] For the meaning see SHADDICK and ZIDEK [2012]

NOTE: $\Delta g(u) = \sum_{j=1}^{d} \frac{\partial^2 g}{\partial u_j^2}$.

To complete the specification, assume

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}}.$$

In our apps, $d = 2$ usually so $\alpha = \nu + 1$. Note that $\nu = 1/2$ gives the exponential variogram.

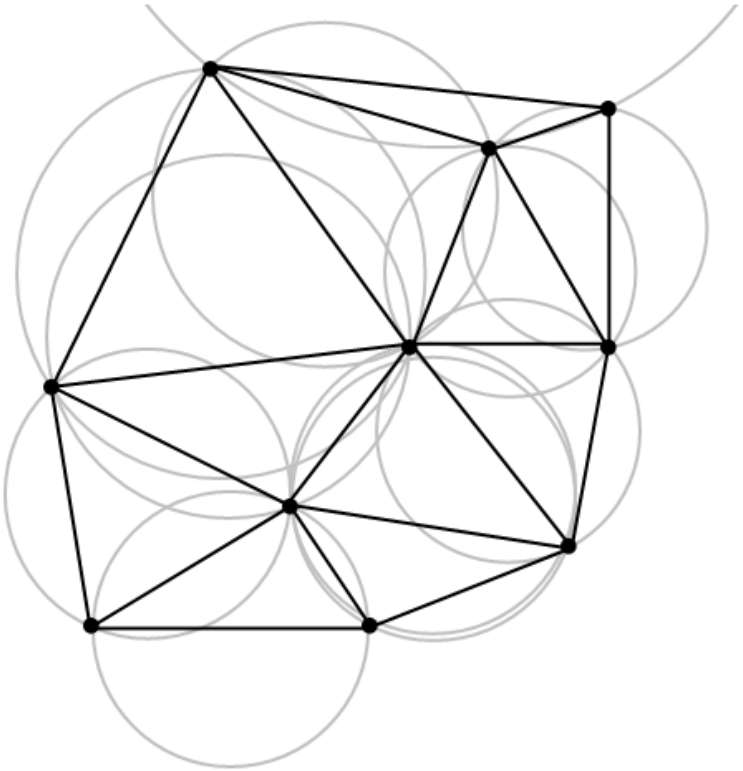
INLA building blocks: the SPDE

Finn Lindgren (personal communication) is writing a book that rebuilds geostatistics on the SPDE approach - replaces Matern with its cousin, the SPDE.

In general approximate numerical solutions to PDEs are required. Done through difference equation approximations or ***finite element approximations***.

The finite elements are a lattice of (Delauney) triangles in R^d on which are built piece wise basis function functions.

Delauney triangulation given monitoring sites. Sites are some of the vertices. The triangular array can be extended by adding some constraints.No new or old site lies inside circumcircle of any $\triangle$



Building approximate solutions to the SPDE

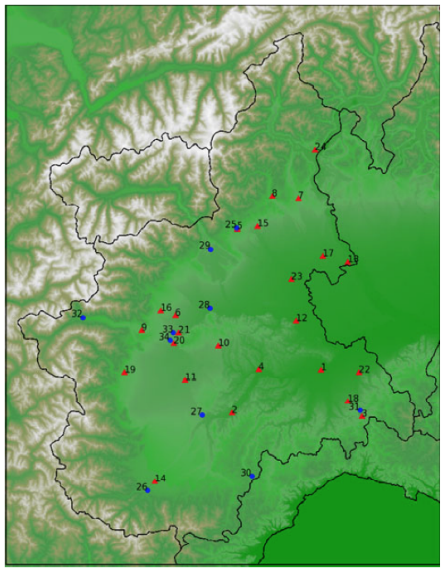
Example[6]: Spatial field Y represents PM_{10} in the Piemonte region.

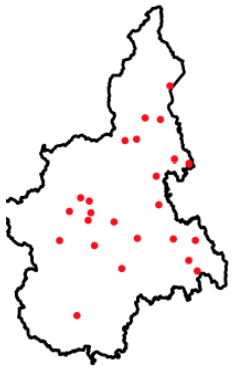
Here is the Piemonte region in Northern Italy.

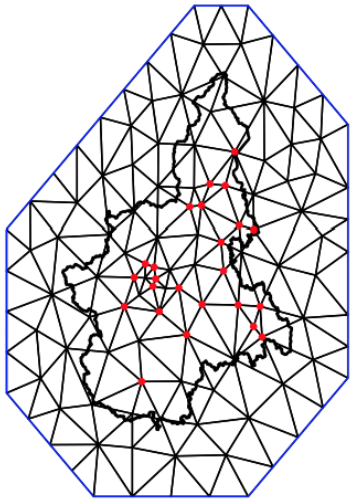
Step 1: Triangulate the region

[6] Cameletti et al. [2011]

Result: Red dots, PM_{10} are monitoring sites. They are grid points of the $\triangle$ mesh lattice.







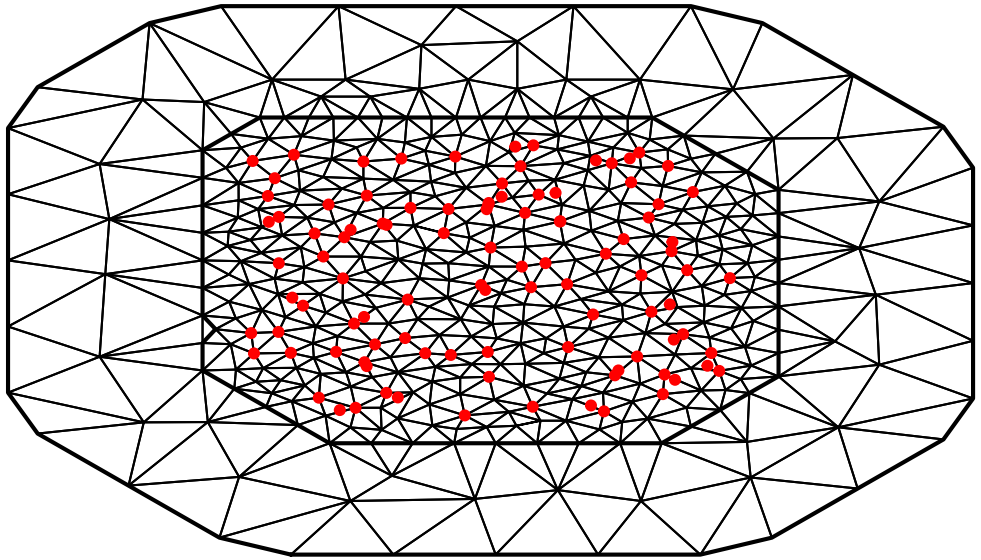
Step 1: Creating the mesh:

```
m = 100
points = matrix(runif(m*2),m,2)
mesh = inla.mesh.create.helper(
  points=points,
  cutoff=0.05,
  offset=c(0.1,0.4),
  max.edge=c(0.05,0.5) )
plot(mesh)
points(points[,1],points[,2])
```

NOTE: Cutoff parameter avoids lots of small $\triangle$s when mesh is extended

Resulting mesh:

Constrained refined Delaunay triangulation



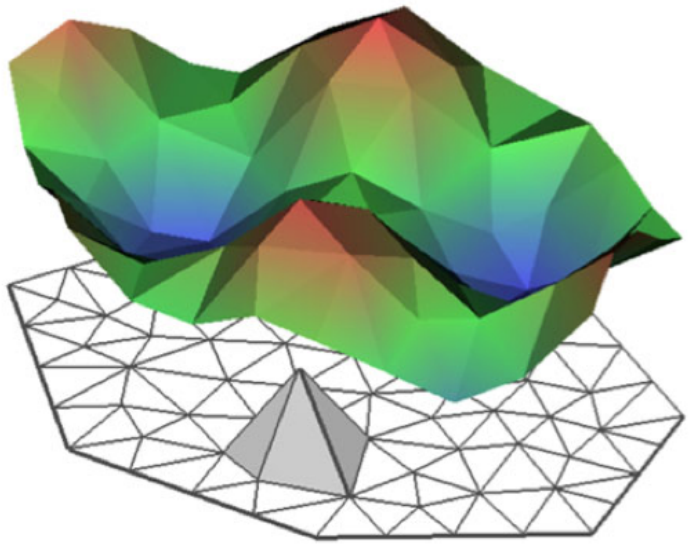
mesh

Next: finite element approximation to solution Y of SPDE:

$$Y^{approx}(s) = \sum_{l=1}^{p} \Psi_l(s)\tilde{\omega}_l^{\text{weights at lattice points}}$$

where $\tilde{\omega} = (\tilde{\omega}_1, \ldots, \tilde{\omega}_p) \sim N_p(0, Q^{-1})$. This links the original Gaussian field (GF) Y, which has Matern covariance, to the GMRF ω through $(\kappa, \nu) \leftrightarrow Q$.

Step 2: Construct "basis functions": Each a piece linear function supported by a $\triangle$. Figure shows one simulation of the field over Piemonte of a finite element approximate solution to SPDE built on a triangular mesh with *p* elements:



Implementing our spatio–temporal model in INLA

Recall the original GF model

$$
\begin{aligned}
Y_t &= (Y^{(g)}, Y^{(u)}) \\
Y^{(g)} &= (Y_t(s_1), \ldots, Y_t(s_g)), \text{ etc.} \\
Z_t &= x_t\beta + Y_t^{(g)} + \epsilon_t,\ \epsilon_t \sim N(0, \sigma_\epsilon^2 I_g) \\
Y_t &= aY_{t-1} + \omega_t,\ \omega_t \sim N(0, \Sigma = \sigma_\omega^2 \tilde{\Sigma}) \\
Y_1 &\sim N(0, \Sigma/(1-a^2))
\end{aligned}
$$

INLA strategy: Replace ω_t by $\tilde{\omega}_t$ from the SPDE linked to the Matern covariance. Moves us from the GF to the GMRF.

Rewritten process model The process is now defined on the lattice starting with the innovations process $\{\tilde{\omega}_t\}$ that replaces $\{\omega_t\}$. We avoid clutter by omitting the "approx" superscript.

$$\begin{aligned} Y_t &= aY_{t-1} + \tilde{\omega}_t,\ \tilde{\omega}_t \sim N(0, Q_S^{-1} \\ Y_1 &\sim N(0, Q_S^{-1}/(1-a^2)) \end{aligned}$$

Assume space – time separability.

Then $Y \sim MVN(0, Q^{-1})$ with $Q = Q_T \otimes Q_S$

Q_T comes from the AR(1) model[7]:

$$\mathbf{Q}_T = \begin{pmatrix} 1/\sigma_\omega^2 & -a/\sigma_\omega^2 & & & \\ -a/\sigma_\omega^2 & (1+a^2)/\sigma_\omega^2 & & & \\ & & \ddots & & \\ & & & (1+a^2)/\sigma_\omega^2 & -a/\sigma_\omega^2 \\ & & & -a/\sigma_\omega^2 & 1/\sigma_\omega^2 \end{pmatrix}$$

[7] Cameletti et al. [2011]

Rewritten data model: Recall that monitors on lattice grid points

$$Z_t(s_i) \quad = \quad x_t^{\text{Covariates}}(s_i)\beta + BY_t(s_i) + \varepsilon_t(s_i)$$

where B "picks off" the relevant process elements from the GMRF approximant.

The original INLA included non Gaussian fields. In the Gaussian case, exact results obtain for the posteriors, i.e. $[y, \beta \mid \theta] \sim N(0, Q^{-1})$ with hyperparameters $\sigma_\omega^2, a, \kappa$. Thus

$$\pi(y, \beta \mid z) \quad \propto \quad \pi(\theta)\pi(y, \beta \mid \theta)\Pi_{t=1}^{T}\pi(z_t \mid y, \beta, \theta)$$

Spatial prediction

INLA moves Y's domain onto the lattice points. Given $Z = z$, Y and hence the responses Z at non-monitored sites are easily predicted.

Appying INLA

Visit the INLA site[8] where numerous examples are given along with tutorials and so on. Includes the program for the Piemonte analysis.

[8] www.r-inla.org

Step 2: Create SPDE Model object: A simple one is:

```
spde=inla.spde2.matern(mesh,alpha=2)
```

But need to incorporate prior information:

```
sigma0 = 1 ## field std.dev
range0 = 0.2
kappa0 = sqrt(8)/range0
tau0 = 1/(sqrt(4*pi)*kappa0*sigma0)
spde=inla.spde2.matern(mesh, B.tau=cbind(log(tau0),1,0),
B.kappa=cbind(log(kappa0),0,1), theta.prior.mean=c(0,0),
theta.prior.prec=1)
```

Now consider simple process model: for each $i = 1, \ldots, m$

$$\begin{aligned} y(s_i) &= \beta_0 + c_i\beta_c + x_1(s_i) + e(s_i) \\ y(s_{i+m}) &= \beta_0 + c_{i+m}\beta_c + x_2(s_i) + e(s_{i+m}) \end{aligned}$$

where c_i is an observation-specific covariate, $e(s_i)$ is measurement noise and x_1 and x_2 are the two field replicates. Note that the offset, β_0, can be interpreted as a spatial covariate effect.

EXERCISE: Work through Finn's tutorial material in the Project 2 website. Also visit the INLA site to see the Examples there.

Back to Piemonte!

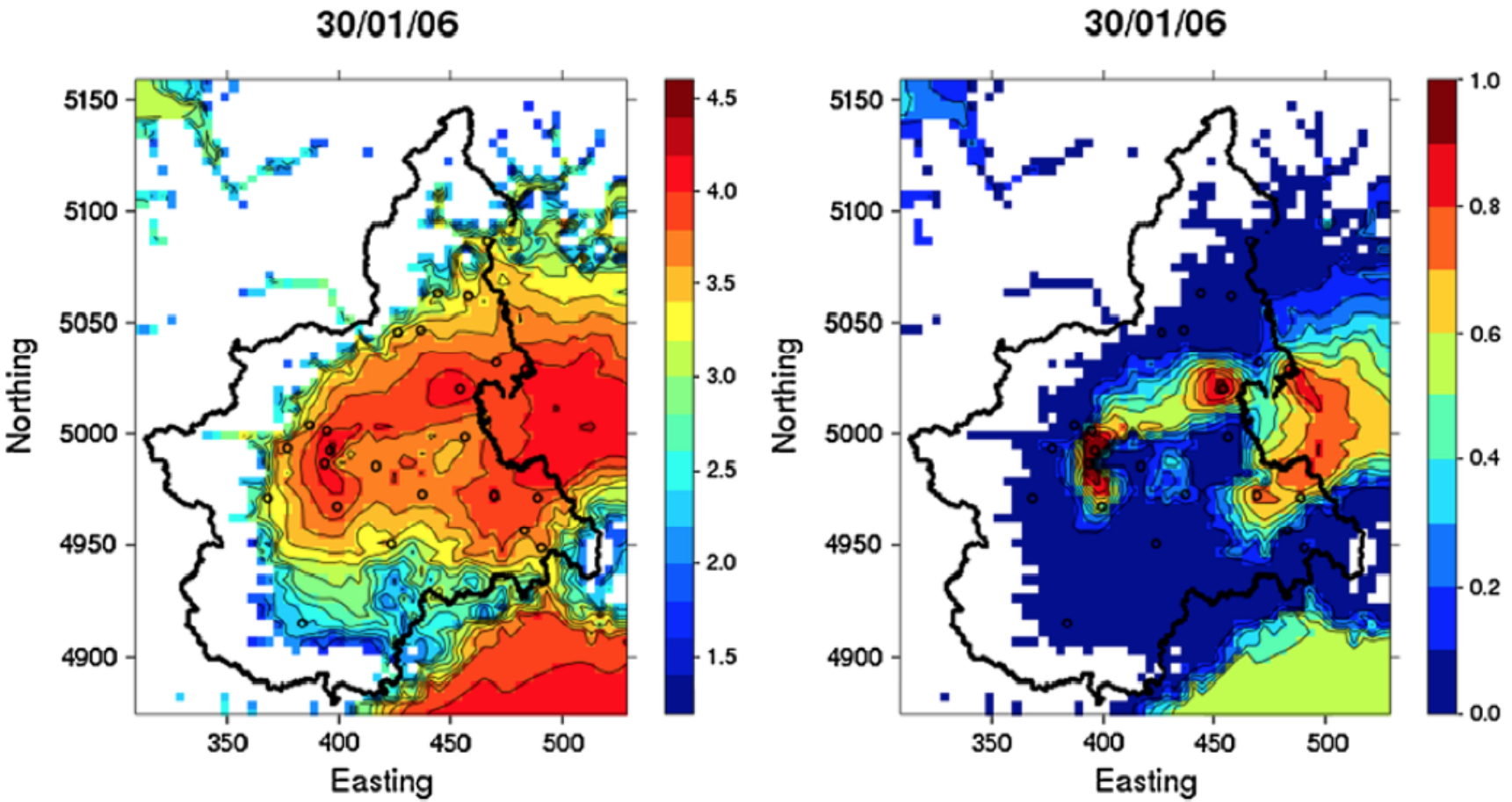


Fig. 5 Map of the PM_{10} posterior mean on the logarithmic scale (*left*) and exceedance probability for 50 $\mu g/m^3$ (*right*) for January 30th, 2006. Only locations with an altitude below 1 km are shown

Summary

(1.) The course has emphasized the link between environmental spatio–temporal and health effects. Temperature will play an increasingly important role.

(2.) Modern software for modeling the relationship has been featured

(3.) The course was “not cookbook” oriented. Skill development has been emphasized including critical and analytical thinking about open ended problems.

(4.) The skills are transferable to professional and academic research environments. Finding the relationship between and Y and X is fundamental.

(5.) Above all understanding uncertainty and modeling its origin, variabilty has been the foundation on which this course was built. But any serious statistics course would also address these issues, albeit in a variety of contexts.

M. Cameletti, F. Lindgren, D. Simpson, and H. Rue. Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, pages 1–23, 2011.

F. Lindgren, H. Rue, and J. Lindstr$\acute{}$om. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

GAVIN SHADDICK and JAMES V ZIDEK. Preferential sampling in long term monitoring of air pollution: a case study. Technical report, Technical Report 267, Department of Statistics, University of British Columbia, 2012.