Spatial models

- Set up
- Classifying spatial data
- Modeling areal data
- Modeling point-referenced data
- Spatial interpolation
Set up

Analysis of spatially distributed data is an area that has been heavily used in practical applications, specially over the last decades:

- Ripley (1981) - 252 pages
- Cressie (1993) - 900 pages

The models start by considering a region with sites or pixels

\[ \{s_1, \ldots, s_d\}. \]

For each site, a variable of interest \( y_i \) is observed with mean \( h(\theta_i) \), for some function \( h \)

\[ E(y_i) = h(\theta_i) \]
Typically, it is assumed that there is an underlying pattern formed by the values of $\theta = (\theta_1, \ldots, \theta_d)'$ associated with the observed scene.

This pattern is corrupted by some random observation mechanism.

The main effort of the inference is to remove this observational noise and recover the underlying, unobserved scene.

Most models commonly assume observational independence conditional on the unobserved image $\theta$, ie.

$$L(\theta) = f(y_1, \ldots, y_d|\theta_1, \ldots, \theta_d) = \prod_i f(y_i|\theta_i)$$

Other effects:

- Explanatory variables - fixed variation;
- Unstructured pixelwise random effects to account for unspecified data heterogeneity.
Classifying spatial data

- **Point-referenced data**
  
  \( y(s) \) is a random vector at location \( s \), which varies continuously over fixed region \( D \).

- **Areal data**
  
  Fixed (regular or irregular) region \( D \) partitioned into a finite number of areal units with well defined boundaries.

- **Point process data**
  
  Random locations in \( D \).
Figure 1.1 from Banerjee et al. (2004) - Map of particulate matter less than 2.5 microns in diameter (PM2.5, measured in ppb) sampling (air pollution monitoring) sites over three midwestern U.S. states: plotting character indicates range of average monitored PM2.5 level over the year 2001.
Point process data

Figure 1.1 from Bognar (2005) - Location of 124 Norway spruce trees in a $56 \times 38$ meter field. Character size is proportional to the trunk diameter.
Areal data

Nobre, Schmidt and Lopes (2005) Spatio-temporal models for mapping the incidence of malaria in Pará, northern Brazil.
US continental states
Adjacency matrix

1 8 9 40 22
2 4 26 42 5 29
3 23 40 22 16 41 34
4 35 26 2
5 48 25 14 34 29 2 42
6 30 19 37
7 28 18 36
8 9 1
9 1 8 38 31 40
10 45 35 26 42 48 24
11 47 12 15 23 13
12 11 20 33 15
13 21 47 11 23 25 39
14 25 23 34 5
15 23 11 12 33 46 44 40
16 41 3 22
17 27
18 36 7 44 46
19 43 27 37 6 30
20 12 33 47
21 32 39 13 47
22 16 3 40 1
23 13 11 15 40 3 34 14 25
24 10 48 39 32
25 39 13 23 14 5 48
26 35 10 42 2 4
27 43 19 17
28 7 30 36
29 2 42 5 34 41
30 43 19 6 28 36
31 44 40 9 38
32 24 39 21
33 20 12 15 46 36
34 14 23 3 41 29 5
35 45 10 26 4
36 30 28 7 16 46 33
37 6 19
38 31 9
39 32 24 48 25 13 21
40 15 44 31 9 1 22 3 23
41 29 34 3 16
42 10 48 5 29 2 26
43 30 19 27
44 46 18 31 40 15
45 35 10
46 33 36 18 44 15
47 20 11 13 21
48 24 39 25 5 42 10
Modeling areal data

The spatial structure is specified through the prior $p(\theta)$.

Besag, York and Mollié (1991) suggest a pairwise difference (PD) prior form

$$p(\theta) \propto \exp \left\{ - \sum_{i<j} w_{ij} h(\theta_i - \theta_j) \right\},$$

where the proportionality constant is not uniquely defined due to its degeneracy.

They consider $h(x) = x^2 / 2W$ leading to a singular normal distribution

$$\theta \sim PD(w, W)$$

where the proportionality constant now includes the term $W^{-d/2}$. 
Local dependence

It is easy to see that, for \( i = 1, \ldots, d \),

\[
\theta_i | \theta_{-i}, W \sim N(a_i, R_i)
\]

where

\[
a_i = \frac{1}{n_i} \sum_{j=1}^{d} w_{ij} \theta_j \quad \text{and} \quad R_i = \frac{1}{n_i} W,
\]

where \( n_i = \sum_{j=1}^{d} w_{ij} \).

\( \theta_i \)'s depend on their spatial neighbours just like in a temporal autoregressive structure.

PD forms are usually referred to as conditional autoregressive (CAR) models.
Choosing $w_{ij}$

A typical choice of weights is

$$w_{ij} = I(s_j \in N_i)$$

where $N_i$ defines a neighbourhood of $s_i$, $i = 1, \ldots, d$.

Other choices of $w_{ij}$ possibly depending on unknown hyperparameters may also be used.

Gamerman and Moreira (2004) considered

$$w_{ij} \propto \frac{1}{d_{ij}^{\tau}}$$

where $d_{ij}$ is some distance between pixels $i$ and $j$ and $\tau$ is an unknown quantity measuring the strength of the spatial dependence.

The distribution above can also be identified with Gaussian Markov random fields (GMRF).
Gaussian random fields (GRF)

A space-varying process $\theta(\cdot)$ taking values in a region $S$ follows a GRF if $\theta = (\theta(s_1), \ldots, \theta(s_d))$ possesses a $d$-variate normal distribution for any $d$ and any given set of $d$ locations $s_1, \ldots, s_d$ in $S$.

The identity $\theta(s_i) = \theta_i$, for all $i$, is frequently used, especially when the locations $s_i$ refer to areas contained in $S$ rather than point locations belonging to $S$.

If the dependence between them is provided by the non-zero weights, a Markovian structure in space is defined.
A collection \( \{ \theta_1, \theta_2, \ldots \} \) is a MRF if the full conditional distributions of \( \theta_i \) depend only on \( \theta_j \) for \( j \in N_i \), the set of neighbors of \( i \), \( i = 1, 2, \ldots \).
Hierarchical structure

A common choice for observation model is given by the normal distribution.

It provides suitable representation (possibly after some transformation) of (i) crop output; (ii) economic indices; or climatic indicators.

- **First level: data structure**
  \[ y | \theta, \sigma^2 \sim N(\theta, \sigma^2 I_d) \]

- **Second level: spatial structure**
  \[ \theta | W \sim CAR(w, W) \]

- **Third level: hyperparameters**
  \[ \sigma^2 \sim IG \left( \frac{n_{\sigma}}{2}, \frac{n_{\sigma} S_{\sigma}}{2} \right) \]
  \[ W \sim IG \left( \frac{n_{W}}{2}, \frac{n_{W} S_{W}}{2} \right) \]
### Full conditionals

It is straightforward to see that

\[
\begin{align*}
\sigma^2 & \sim \text{IG}(n^*_\sigma/2, n^*_\sigma S^*_\sigma/2) \\
W & \sim \text{IG}(n^*_W/2, n^*_W S^*_W/2) \\
\theta_i & \sim N(m_i, C_i)
\end{align*}
\]

for \(i = 1, \ldots, d\), with

\[
\begin{align*}
n^*_\sigma & = n_\sigma + d \\
n^*_\sigma S^*_\sigma & = n_\sigma S_\sigma + \sum (y_i - \theta_i)^2 \\
n^*_W & = n_W + d \\
n^*_W S^*_W & = n_W S_W + \sum w_{ij}(\theta_i - \theta_j)^2 \\
m_i & = C_i(\sigma^{-2}y_i + W^{-1} \sum w_{ij}\theta_j) \\
C_i^{-1} & = \sigma^{-2} + n_i W^{-1}
\end{align*}
\]

for \(i = 1, \ldots, d\).
Computational issues

In practical applications the value of $d$ is usually large, making direct sampling of $\theta$ very slow.

Rue (2001) suggested numerical techniques to take advantage of the sparseness (presence of many 0s) of $C^{-1}$ ($\theta$’s precision matrix) and hence improve sampling considerably.

These techniques basically involve appropriate reordering of the pixels to ensure a minimal diagonal band structure to $C^{-1}$ followed by a numerically efficient Cholesky decomposition that takes advantage of this band diagonalization.

See Gamerman, Moreira and Rue (2003) and Knorr-Held and Rue (2002) for further discussion on blocking advantages and disadvantages.
Figure 1. (a) The map of Germany with $n = 544$ regions, and (b) the corresponding graph for the GMRF where neighbour regions share a common border.

Figure 2. (a) The black regions make north and south conditional independent, and (b) displays the automatic computed reordering starting from the white region ending at the black region. This reordering produce the precision matrix in Figure 3b.

Figure 3. (a) The precision matrix $Q$ is the original ordering, and (b) the precision matrix after reordering to facilitate a band matrix with small bandwidth.

© 2003 Håvard Rue http://www.math.ntnu.no/~hrue
Extensions

Higdon and Besag (1999) considered the inclusion of a regression term $X\beta$ in the mean structure of the observations:

$$ y|\theta, \sigma^2 \sim N(X\beta + \theta, \sigma^2 I_d) ; $$

Congdon (1997) introduced the idea of allowing spatial variation also to regression coefficients $\beta$;

Assunção, Gamerman and Assunção (1999) used the multivariate form

$$ p(\beta) \propto \exp \left\{ - \sum_{i<j} w_{ij} (\beta_i - \beta_j)' W^{-1} (\beta_i - \beta_j) \right\} $$

for jointly modelling the prior for space-varying regression coefficients $\beta = (\beta_1, \ldots, \beta_d)$.

Observations in the exponential family of distributions (Banerjee, Carlin and Gelfand, 2004, and Rue and Held, 2005).
Modeling point-referenced data

Under the usual simplifying assumptions of common mean, homoscedasticity (common variance) and isotropy (correlation depends only on distance),

$$ \theta \sim N(\mu 1_d, \tau^2 R_\lambda) $$

where

$$ R_\lambda = (\rho_{ij}) $$

with

$$ \rho_{ij} = \rho_\lambda(|s_i - s_j|) $$

for some suitably defined correlation function $\rho_\lambda$ possibly depending on a parameter $\lambda$, typically a scalar or low dimensional quantity.

This distribution is denoted by

$$ \theta(\cdot) \sim GRF(\mu, \tau^2, \rho_\lambda) $$
Correlation functions

The most used correlation functions are

(i) the exponential

$$\rho_\lambda(u) = e^{-u/\lambda},$$

(ii) the Gaussian

$$\rho_\lambda(u) = e^{-u^2/\lambda^2} \text{ and}$$

(iii) the Matérn

$$\rho_{\lambda,\nu}(u) = \frac{1}{2^{\nu-1} \Gamma(\nu)} (2u\lambda\sqrt{\nu})^\nu B_\nu(2u\lambda\sqrt{\nu}).$$

\[\nu = 1/2 \Rightarrow \text{Exponential}\]
\[\nu \rightarrow \infty \Rightarrow \text{Gaussian}\]

• Stein (1999) and Banerjee et al. (2004) recommend the family of Matérn correlation functions.
This prior is based on the notion of distance defining the space. It is better suited for \textit{spatially continuous} data where observations are measured on a specific point location and pixels are associated with points in the space of interest.

Nevertheless, there is nothing compulsory about these guidelines. \textbf{GMRF may also be used with spatially continuous data} (after appropriate definition of neighbourhood of points) just like

\textbf{Distance-based GRF may be used with spatially discrete data} (after appropriate definition of distance between areas).
Hierarchical model

- **First level: data structure**
  \[
  y | \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2 I_d)
  \]

- **Second level: spatial structure**
  \[
  \theta | \mu, \tau^2, \lambda \sim \text{GRF}(\mu, \tau^2, \rho_{\lambda})
  \]

- **Third level: hyperparameters**
  \[
  \begin{align*}
  \mu &\sim \mathcal{N}(a, R) \\
  \sigma^2 &\sim \text{IG}(n_\sigma/2, n_\sigma S_\sigma/2) \\
  \tau^2 &\sim \text{IG}(n_\tau/2, n_\tau S_\tau/2) \\
  \lambda &\sim p(\lambda)
  \end{align*}
  \]

The prior for \( \lambda \) depends on the choice of correlation function (\( \lambda \) can be a vector).
Full conditionals

- $\theta | \Psi \sim N(m_\theta, C_\theta)$
  
  \[
  m_\theta = C_\theta (\tau^{-2} R_\lambda^{-1} 1_d \mu + \phi y) \\
  C_{\theta}^{-1} = \tau^{-2} R_\lambda^{-1} + \phi I_d
  \]

- $\mu | \theta, \phi, \tau^2, \lambda \sim N(m_\mu, C_\mu)$
  
  \[
  m_\mu = C_\mu (\tau^{-2} 1_d R_\lambda^{-1} \theta + R^{-1} a) \\
  C_\mu = (\tau^{-2} 1_d R_\lambda^{-1} 1_d + R^{-1})
  \]

- $\phi | \theta, \mu, \tau^2, \lambda \sim G(n_\sigma^*/2, n_\sigma^* S_\sigma^*/2)$
  
  \[
  n_\sigma^* = n_\sigma + d \\
  n_\sigma^* S_\sigma^* = n_\sigma S_\sigma + \sum [y_i - \theta(s_i)]^2
  \]

- $\tau^2 | \theta, \mu, \phi, \lambda \sim IG(n_\tau^*/2, n_\tau^* S_\tau^*/2)$
  
  \[
  n_\tau^* = n_\tau + d \\
  n_\tau^* S_\tau^* = n_\tau S_\tau + Q_\lambda \\
  Q_\lambda = (\theta - 1_d \mu)' R_\lambda^{-1} (\theta - 1_d \mu)
  \]

- $\lambda | \theta, \mu, \phi, \tau^2 \sim F_\lambda$ with density
  
  \[
  \pi(\lambda | \theta, \mu, \phi, \tau^2) \propto p(\lambda) |R_\lambda|^{-1/2} \exp\{-\frac{1}{2\tau^2} Q_\lambda\}.
  \]
Computational issues

A major computational problem arises here when $d$ is large.

The inversion of the square matrix $R_{\lambda}$ of order $d$ is required at every iteration, slowing down the computations.

Nevertheless, it is expected in the case of large $d$ to have many (possibly most) entries in $R_{\lambda}$ with negligible values due to many small correlation values.

This fact can be used to design approximation methods to band diagonalize this matrix and ensure application of fast inversion algorithms (see Rue and Tjelmeland (2004) for more details).
Spatial interpolation

GRF provide a flexible tool to allow for prediction to unobserved locations or spatial interpolation based on a set of observed location.

This procedure is commonly referred to by geologists as kriging, named after D.G.Krige, a South African mining engineer who worked on geostatistical data.

Bayesian kriging: spatial interpolation based on observed data

\[
  f(y^{(u)}|y^{(obs)}) = \int p(y^{(u)}, \theta^{(u)}, \theta^{(obs)}, \Psi|y^{(obs)}) d\theta^{(u)} d\theta^{(obs)} d\Psi \\
  = \int f(y^{(u)}|\theta^{(u)}, \theta^{(obs)}, \Psi, y^{(obs)}) \times p(\theta^{(u)}|\theta^{(obs)}, \Psi, y^{(obs)}) \times p(\theta^{(obs)}, \Psi|y^{(obs)}) d\theta^{(u)} d\theta^{(obs)} d\Psi \\
  = \int f(y^{(u)}|\theta^{(u)}, \Psi) p(\theta^{(u)}|\theta^{(obs)}, \Psi) \times p(\theta^{(obs)}, \Psi|y^{(obs)}) d\theta^{(u)} d\theta^{(obs)} d\Psi .
\]
Approximating $f(y(u)|y^{(obs)})$ by MCMC

1. Let $(\theta_1^{(obs)}, \Psi_1), \ldots, (\theta_M^{(obs)}, \Psi_M)$ be a sample from the posterior $p(\theta^{(obs)}, \Psi|y^{(obs)})$;

2. Sample $\theta_i^{(u)}$ from $p(\theta^{(u)}|\theta_i^{(obs)}, \Psi_i)$, for $i = 1, \ldots, M$;

3. Compute

$$\hat{f}(y(u)|y^{(obs)}) = \frac{1}{M} \sum_{i=1}^{M} f(y(u)|\theta_i^{(u)}, \Psi_i)$$
Extensions

- Universal kriging: inclusion of $X\beta$;
- Exponential family: Diggle et al. (1998);
- Spatial variation to $\beta$: Gelfand et al. (2003);
- Dynamic spatio-temporal data: Gelfand et al. (2005).


