Spatial Reading Group

February 16, 2017

Outline

Motivation

Why Spatial Data needs Spatial Stats

- Spatial Data are continuous but measured discretely.
- As a result the measurements tend to be correlated.
- The measurements are rarely taken at random
- Different versions of distance.

Figure: *The Geevor Mine*

Outline

Assumptions

For the rest of the presentation we are in $\mathbb{R}^2$

- Smooth data (Except for "nugget effect")
- Stationary data
  - Does a trend extend beyond the bounds of the study?
  - Is the covariance consistent in the bounds of the study?
- $\{S(x) : x \in \mathbb{R}^2\}$ Is Gaussian with
  - mean $= \mu$
  - variance $\sigma^2 = Var\{S(x)\}$
  - correlation fxn $\rho(u) = Corr\{S(x), S(x')\}$
  - $u_i = ||x_i - x'||$
- $Y_i = S(x_i) + Z_i$

Spatial Covariance $\rho(u)$

- Similar to Autocovariance in time series, except time lag is replaced by distance u.

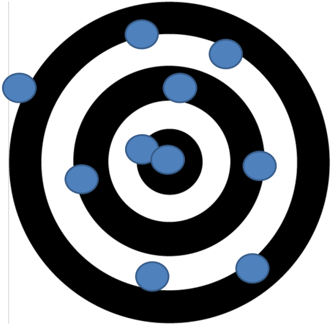


Figure: *Measurements are some distance u from each other.*

Mattern Function

- General function describing how quickly the correlation decays.
- $\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1} \frac{u}{\phi}^{\kappa} K_{\kappa}(\frac{u}{\phi})$
  - κ: order of differentiation, smoothness.
  - ϕ: scale, degree of decay over time.
  - $K_{\kappa}()$: Modified Bessel function.
- Special cases:
  - $\kappa = 0.5$: Exponential decay in $\mathbb{R}^2$.
  - $\kappa \to \infty$: Gaussian decay in $\mathbb{R}^2$.
- κ and ϕ are not orthogonal.

Outline

Variogram

Visualising the decay of $\rho(u)$ with distance

- Bin the variance between all the points into set distances
- τ^2: The nugget
- $\tau^2 + \sigma^2 = Var(y)$: The sill
- When $V_Y(u) = Var(y)$: Range

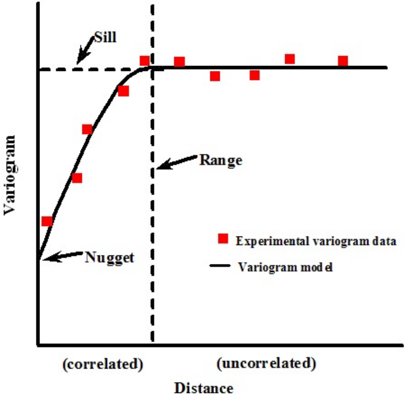


Figure: *Estimating how correlation changes with distance.*

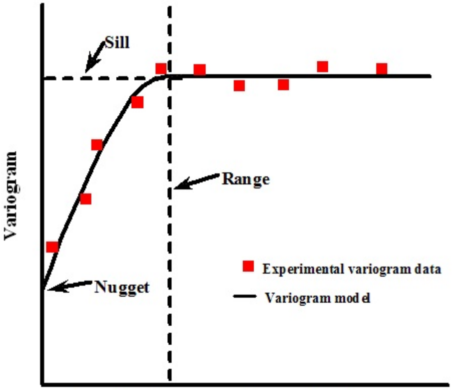
Variogram part 2

What is it actually?

$$V(u) = \frac{1}{2} Var\{S(x)S(xu)\} \tag{1}$$

$$= \sigma^2\{1 - \rho(u)\} \tag{2}$$

$$V_Y(u) = \tau^2 + \sigma^2\{1 - \rho(u)\} \tag{3}$$



Maximum Likelihood Estimation

- Gaussian model

$$Y \sim N(D\beta, \sigma^2 R(\phi) + \tau^2 I)$$

with covariates matrix $D_{n \times p}$, regression coefficients β , covariance of a parametric model for $S(x)$, and nugget variance τ^2.

- The log-likelihood function is

$$\begin{aligned} L(\beta, \tau^2, \sigma^2, \phi) = & -0.5\{n \log(2\pi) + log\{|(\sigma^2 R(\phi) + \tau^2 I)|\} \\ & + (y - D\beta)^T (\sigma^2 R(\phi) + \tau^2 I)^{-1} (y - D\beta)\} \end{aligned}$$

Maximum Likelihood Estimation

- Let $\nu^2 = (\tau^2/\sigma^2)$, $V = R(\phi) + \nu^2 I$, then $L(\beta, \tau^2, \sigma^2, \phi)$ is maximized at

$$\hat{\beta}(V) = (D^T V^{-1} D)^{-1} D^T V^{-1} y \tag{4}$$
$$\hat{\sigma}^2(V) = n^{-1}\{y - D\hat{\beta}(V)\}^T V^{-1}\{y - D\hat{\beta}(V)\} \tag{5}$$

- Plug (1) and (2) into $L(\beta, \tau^2, \sigma^2, \phi)$ and obtain the concentrated log-likelihood:
  $L_0(\nu^2, \phi) = -0.5\{n \log(2\pi) + n \log \hat{\sigma}^2(V) + log|V| + n\}$
- Optimize $L_0(\nu^2, \phi)$ numerically with respect to ν and ϕ; back substitution to obtain $\hat{\sigma}^2$ and $\hat{\beta}$

Maximum Likelihood Estimation

- Re-parameterisation of V can be used to obtain more stable estimation, e.g the ratio σ^2/ϕ is more stable than σ^2 and ϕ
- Computational tool: profile log-likelihood:
  Assume a model with parameters (α, ψ),

$$L_p(\alpha) = L(\alpha, \hat{\psi}(\alpha)) = \max_{\psi}(L(\alpha, \psi))$$

Maximum Likelihood Estimation

- Non-Gaussian data:
  (1): transformation to Gaussian (2) generalized linear model
- (1) E.g. Box-Cox transformation; denote the transformed responses $Y^* = (Y_1^*, ..., Y_n^*)$, and fit a Gaussian model

$$Y^* \sim N(D\beta, \sigma^2\{R(\phi) + \nu^2 I\})$$

  Computationally demanding, transformation may impede scientific interpretation
- (2) Generalized linear model

$$L(\theta|S) = \prod_{i=1}^{n} f_i(y_i|S, \theta)$$

$$L(\theta, \phi) = \int_S \prod_{i=1}^{n} f_i(y_i|s, \theta) g(s|\phi) ds$$

  Involve high dimensional integration;need MCMC/Hierarchical likelihood/Generalized estimating equations

Maximum Likelihood Estimation (An Example)

Model with constant mean							
Model	$\hat{\mu}$			$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	logL
$\kappa = 0.5$	863.71			4087.6	6.12	0	−244.6
$\kappa = 1.5$	848.32			3510.1	1.2	48.16	−242.1
$\kappa = 2.5$	844.63			3206.9	0.74	70.82	−242.33
Model with linear trend							
Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	logL
$\kappa = 0.5$	919.1	−5.58	−15.52	1731.8	2.49	0	−242.71
$\kappa = 1.5$	912.49	−4.99	−16.46	1693.1	0.81	34.9	−240.08
$\kappa = 2.5$	912.14	−4.81	−17.11	1595.1	0.54	54.72	−239.75

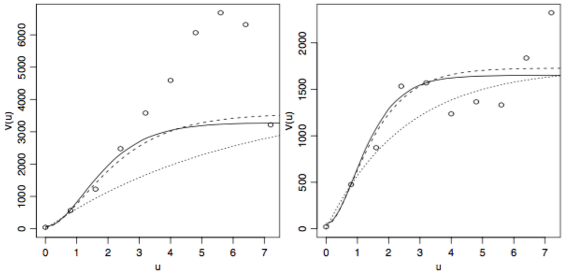


Figure: *left: constant mean model; right: linear trend surface*

circle: sample variogram; solid line $\kappa = 2.5$; *dashed line* $\kappa = 1.5$; *dotted line* $\kappa = 0.5$

Kriging

- Suppose our objective is to predict the value of the signal at an arbitrary location $S(x)$.
- Note that $(S(x), Y)$ is multivariate Gaussian with mean vector $\mu\mathbf{1}$ and covariance matrix

$$\begin{pmatrix} \sigma^2 & \sigma^2 r^T \\ \sigma^2 r & \sigma^2 V \end{pmatrix},$$

where r is a vector with elements $r_i = \rho(\|x - x_i\|)$ and $V = \sigma^2 R + \tau^2 I$.

Kriging

- Conditional mean and variance:

$$E(S(x)|Y) = \mu + r^T V^{-1}(Y - \mu\mathbf{1}),$$

$$Var(S(x)|Y) = \sigma^2(1 - r^T V^{-1} r).$$

- Two types:
  - Ordinary kriging: replace μ by its weighted least squares estimator
  $$\hat{\mu} = (\mathbf{1}^T V^{-1}\mathbf{1})^{-1}\mathbf{1}^T V^{-1} Y.$$
  - Simple kriging: replace μ by $\hat{\mu} = \bar{y}$.
- Both kriging predictors can be expressed as a linear combination: $\hat{S}(x) = \sum_{i=1}^n a_i(x) Y_i$, but $\sum_{i=1}^n a_i(x) = 1$ only for ordinary kriging.

Preferential Sampling

The Problem

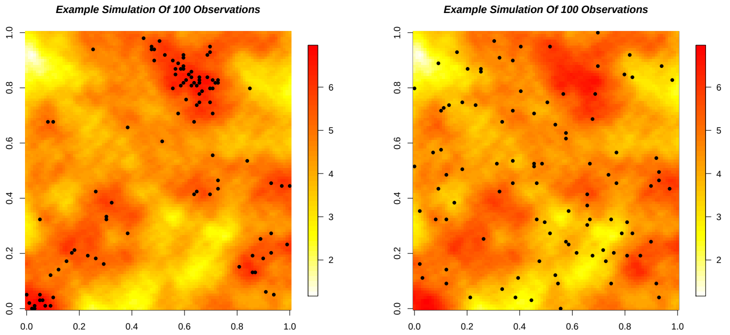
- So far we have assumed the sampling locations X are fixed, or assumed known.
- What if the sampling locations depend on the underlying field S?

Example

- Pollution data from measuring stations
- Ocean temperature data from marine mammals

Preferential Sampling

Figure: Example of a single realisation of S and corresponding 100 sampling locations selected using a spatial Poisson Process with intensity $\lambda(x) = \exp(\beta S(x))$.



(a) Example of 100 preferentially sampled locations ($\beta = 2$)

(b) Example of 100 non-preferentially sampled locations ($\beta = 0$)

Preferential Sampling

Solution

- We must account for the dependence between X and S.

$$L(\boldsymbol{\theta}) = \int [X, Y, S] \, \mathrm{d}S. \tag{6}$$

- Diggle et al. 2010 - Monte Carlo
- Integrated Nested Laplace Approximation (INLA) - Joe
- Template Model Builder - Danny

Preferential Sampling

Results

Model	Parameter	Standard MLE	*TMB*
Preferential	Bias	(0.77, 1.36)	(0.41, 0.94)
Preferential	Root-mean-square error	(0.86, 1.40)	(0.60, 1.05)

Table: Comparison of approximate 95% confidence intervals for the root-mean-square errors and bias between standard MLE and *TMB* over 50 independent simulations for preferential ($\beta = 2$) at location $x_0 = (0.49, 0.49)$.

Problems with the MLE approach

- MLE method seperates parameter estimation and spatial prediction as two distinct problems.
- First the model is formulated, and it's parameters estimated.
- These estimated parameters are assumed true and spatial prediction equations are computed with these estimates plugged-in.
- Parameter uncertainty is ignored when making spatial prediction.
- Parameter uncertainty is often VERY HIGH. Even with seemingly large datasets ($n > 10,000$), the positive correlation dilutes the information present. Largely different values of correlation parameters ϕ often fit the data equally well.

Bayesian approach

- Account for parameter uncertainty when making spatial prediction and hence make more conservative estimates of prediction accuracy.

$$[S|Y] = \int [S|Y,\theta][\theta|Y]d\theta$$

- The Bayesian predictive distribution is a weighted average of plug-in predictive distributions $[S|Y,\hat{\theta}]$, weighted by the posterior uncertainty of the model values θ.
- Arbitrary nonlinear functional $T(S)$ of S can be estimated (along with credible intervals, standard errors etc) by simple deterministic transformations of the posterior samples of S.

Problems with Bayesian implementation

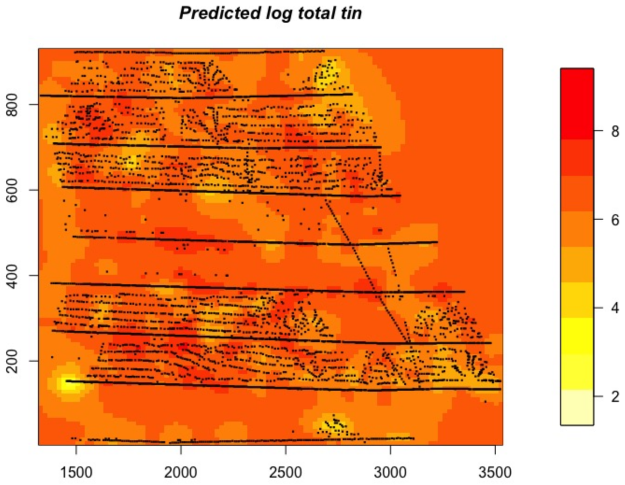
- Due to the flexibility of the Matern correlation function, many different combinations of the correlation parameters ϕ fit the data equally well.
- A consequence of this is that the posterior distribution of $[\theta|Y]$ has non-negligible probability mass across a wide range of the parameter space.
- The first consequence of this is that posterior distributions are extremely sensitive to prior distributions. Apparently 'diffuse' priors can still heavily affect the location and scale of the posterior distribution.
- Secondly, MCMC samplers must be formulated with large transition jumps to ensure the whole parameter space is explored. This leads to VERY LONG MCMC chains needing to be run (100,000 $+$).

The joy of INLA

- INLA enables **very** accurate deterministic approximations to both $[\theta|Y]$ and $\int[S|Y,\theta][\theta|Y]d\theta$ to be obtained.
- INLA handles most well-known response functions (Binomial, Poisson, Gamma etc), enabling Generalized Geostatistical Models to be fit.
- Through a combination of high-accuracy Laplace approximations and cubic spline interpolation, values from INLA are often indistinguishable from the 'true' values from an MCMC chain.
- INLA is FAST. Taking only seconds - minutes to run compared with the hours - days that MCMC can take.
- Multiple responses can be fit to the same spatial process! (E.g poisson process and intensity process enabling preferential sampling to be investigated.)

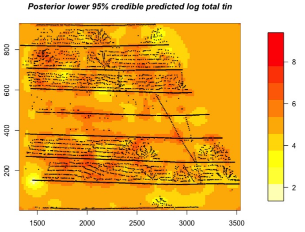
Real example: Predicting total Tin in Cornwall, UK

Figure: Predicted mean field .

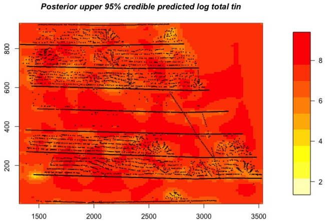


Real example: Predicting total Tin in Cornwall, UK

Figure: Upper and lower 95% credible fields.



(a)



(b)