Univariate spatial modeling and data analysis

Abhi Datta

04.17.2017

Review of last lecture

- Variogram models – Intrinsic stationarity, Fitting a model based variogram to the empirical variograms
- Ordinary kriging – based on variograms
- Covariance functions – weak stationarity, relationship with variograms, isotropy
- Gaussian Processes – likelihood, kriging
- Spatial linear regression

$$y(\mathbf{s}_i) = x(\mathbf{s}_i)\beta + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i); w(s) \sim GP(0, C(\cdot, \cdot \,|\, \theta)$$

Revisiting the exponential covariance function

$$C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t = 0 \\ \sigma^2 \exp(-\phi t) & \text{if } t > 0 \end{cases} .$$

- When range $= \infty$, we use the more informative *effective range*, t_0, defined as the distance at which this correlation has dropped to only 0.05.
- For exponential model, effective range $\approx 3/\phi$.
- Nugget τ^2 is often viewed as a "nonspatial effect variance," and the partial sill (σ^2) is viewed as a "spatial effect variance."
- Note discontinuity at 0 due to the nugget. Intentional! To account for measurement error or micro-sale variability.

Univariate spatial analysis

- Data $\{y(s), x(s)\}$ observed at n locations $s_1, s_2, \ldots, s_n$
- Spatial linear regression model:

$$y(\mathbf{s}_i) = x(\mathbf{s}_i)'\beta + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i); w(s) \sim GP(0, C(\cdot, \cdot \,|\, \theta)$$

Univariate spatial analysis

- Data $\{y(s), x(s)\}$ observed at n locations $s_1, s_2, \ldots, s_n$
- Spatial linear regression model:

$$y(\mathbf{s}_i) = x(\mathbf{s}_i)'\beta + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i); w(s) \sim GP(0, C(\cdot, \cdot \,|\, \theta)$$

- As $\epsilon(s) \overset{\text{iid}}{\sim} N(0, \tau^2)$, $w(s) + \epsilon(s) \sim GP(0, C_1(\cdot, \cdot \,|\, \theta))$ where $C_1(s_i, s_j \,|\, \theta, \tau^2) = C(s_i, s_j \,|\, \theta) + \tau^2 I(s_i = s_j)$
- The function C is chosen without a nugget as $\epsilon(s)$ accounts for the nugget
- Marginalized model: $y \,|\, \theta, \tau^2 = N(X\beta, \Sigma(\theta, \tau^2))$ where $\Sigma = (C(s_i, s_j \,|\, \theta)) + \tau^2 I$
    - Estimate $\hat{\theta}$ and $\hat{\tau^2}$ by ML or REML
    - $\hat{\beta} = (X'\Sigma(\hat{\theta}, \hat{\tau^2})^{-1} X)^{-1} X'\Sigma(\hat{\theta}, \hat{\tau^2})^{-1} y$

Kriging

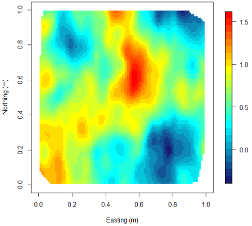
- For any new s_0,
  $Cov(y(s_0), y) = c_0(\theta) = (C(s_1, s_0 \,|\, \theta), \ldots, C(s_n, s_0 \,|\, \theta))'$
- Kriging: $y(s_0) \,|\, y \sim N(x'\hat{\beta} + c_0(\hat{\theta})' \, \Sigma(\hat{\theta}, \hat{\tau^2})^{-1} \, (y - X\hat{\beta}), \hat{\sigma^2} + \hat{\tau^2} - c_0(\hat{\theta})' \, \Sigma(\hat{\theta}, \hat{\tau^2})^{-1} \, c_0(\hat{\theta}))$

Model comparison

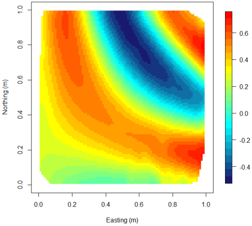
- Model based approaches:
  - AIC: $2k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau^2}))$
  - BIC: $\log(n)k - 2\log(l(y \mid \hat{\beta}, \hat{\theta}, \hat{\tau^2}))$
- Prediction based approaches using holdout data:
  - Root Mean Square Predictive Error (RMSPE): $\sqrt{\frac{1}{n_{out}} \sum_{i=1}^{n_{out}} (y_i - \hat{y}_i)^2}$
  - K-fold Cross-validation based RMSPE: $\sqrt{\frac{1}{K} \sum_{k=1}^{K} \frac{1}{|fold_k|} \sum_{i \in fold_k} (y_i - \hat{y}_i)^2}$
  - Coverage probability: $\frac{1}{n_{out}} \sum_{i=1}^{n_{out}} I(y_i \in (\hat{y}_{i,0.025}, \hat{y}_{i,0.975}))$
  - Width of 95% confidence interval: $\frac{1}{n_{out}} \sum_{i=1}^{n_{out}} \hat{y}_{i,0.975} - \hat{y}_{i,0.025}$
  - The last two approaches compares the distribution of y_i instead of comparing just their point predictions

Data analysis

- Dataset 3 from Lecture 1



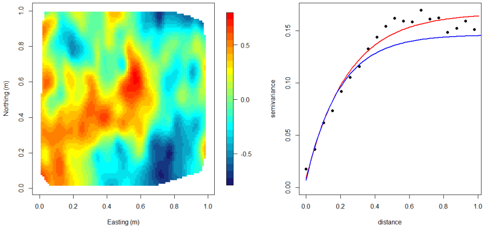
(a) Data 3



(b) $x(s)$

Residual plots

- Linear regression: $y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + \epsilon(\mathbf{s}_i)$



(a) Residual surface

(b) Empirical and fitted variograms. Red: Least squares fit, Blue: MLE

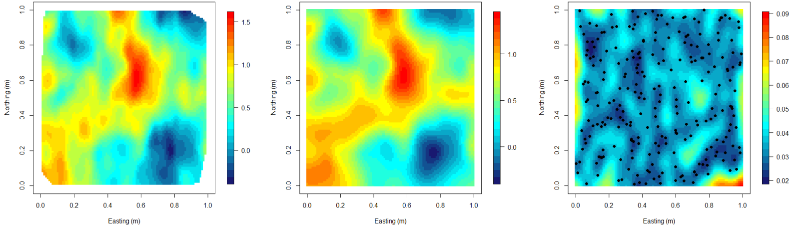
- Spatial model: $y(\mathbf{s}_i) = \beta_0 + x(\mathbf{s}_i)\beta_1 + \epsilon(\mathbf{s}_i)$
- $w(s)$ modeled as an exponential GP

Model comparisons

Model	Non-spatial	Spatial
AIC	224	-32
BIC	235	-15
RMSPE	0.34	0.17
CP	96.4%	96%
CI width	1.47	0.72

- Spatial model performs better

Kriged surfaces



(a) Data 3 (b) Kriged surface (c) Kriging variances

Real data analysis: Working with Latitudes and longitudes

- Most point referenced data are observed on the surface of the earth
- Co-ordinates are reported in longitude (λ) and latitude (θ)
- Can we use (λ, θ) as the locations in our GP models?

Real data analysis: Working with Latitudes and longitudes

- Most point referenced data are observed on the surface of the earth
- Co-ordinates are reported in longitude (λ) and latitude (θ)
- Can we use (λ, θ) as the locations in our GP models?
- The earth is round! So (longitude, latitude) $\neq (x, y)$!
- Using (λ, θ) will heavily distort the distances especially if the data region is even moderately large
- Another approach would be to embed the sphere in $\mathbb{R}^3$ and use the Euclidean distances
- A more natural choice of distance would be the shortest path along the surface of the earth – geodesic distance

Calculating geodesic distances

- From elementary trigonometry, the coords on a sphere are

$$x = R\cos\theta\cos\lambda,\ y = R\cos\theta\sin\lambda, \ \text{ and } \ z = R\sin\theta$$

- Assume a unit sphere (i.e. $R = 1$). Letting $\mathbf{u}_1 = (x_1, y_1, z_1)$ and $\mathbf{u}_2 = (x_2, y_2, z_2)$, we know

$$\cos\phi = \frac{\langle \mathbf{u}_1, \mathbf{u}_2\rangle}{||\mathbf{u}_1||\,||\mathbf{u}_2||} = \langle \mathbf{u}_1, \mathbf{u}_2\rangle\ .$$

- We now compute

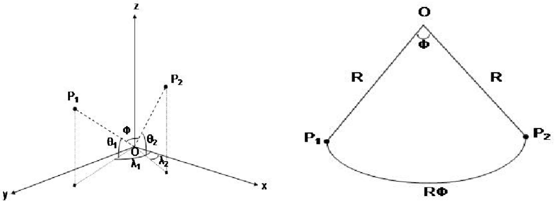
$$\begin{aligned}\langle \mathbf{u}_1, \mathbf{u}_2\rangle &= \cos\theta_1\cos\lambda_1\cos\theta_2\cos\lambda_2 + \cos\theta_1\sin\lambda_1\cos\theta_2\sin\lambda_2 \\ &\quad + \sin\theta_1\sin\theta_2 \\ &= \cos\theta_1\cos\theta_2\cos(\lambda_1 - \lambda_2) + \sin\theta_1\sin\theta_2\end{aligned}$$

- For a sphere of radius R, our final answer is

$$D = R\phi = R\arccos[\cos\theta_1\cos\theta_2\cos(\lambda_1 - \lambda_2) + \sin\theta_1\sin\theta_2]\ .$$

Geodesic distance

The basic geometry behind calculating geodesic distances



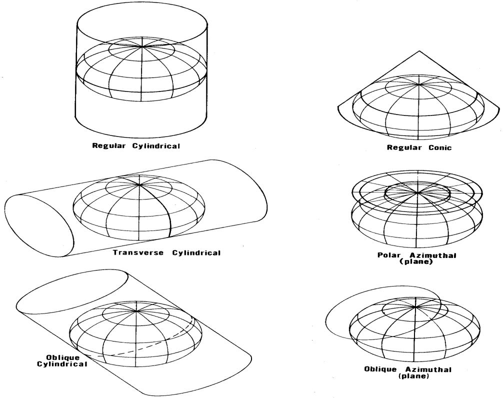
- Consider two points on the surface of the earth, $P_1 = (\theta_1, \lambda_1)$ and $P_2 = (\theta_2, \lambda_2)$, where $\theta =$ latitude and $\lambda =$ longitude.
- The geodesic distance we seek is $D = R\phi$, where
  - R is the radius of the earth
  - ϕ is the angle subtended by the arc connecting P_1 and P_2 at the center

Fundamentals of Cartography

- A map projection is a systematic representation of all or part of the surface of the earth on a plane.
- *Theorem:* The sphere cannot be flattened onto a plane without distortion
- Instead, use an intermediate surface that can be flattened. The sphere is first projected onto the this developable surface as $(f(\lambda, \phi), g(\lambda, \phi))$, which is then laid out as a plane.
- The three most commonly used surfaces are the cylinder, the cone, and the plane itself. Using different orientations of these surfaces lead to different classes of map projections...

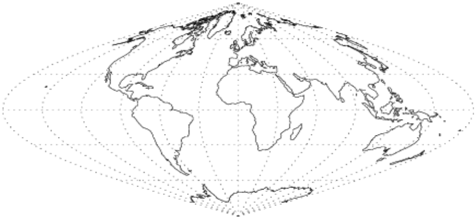
Developable surfaces

Geometric constructions of projections



Cylindrical (Sinusoidal) projection

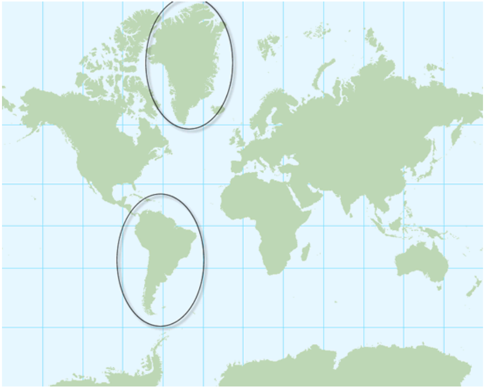
Cylindrical (Sinusoidal) projection



This *sinusoidal* projection obtained by specifying $\partial g/\partial \phi = R$, which yields equally-spaced straight lines for the parallels, and results in (with the 0 degree meridian as the central meridian),

$$f(\lambda, \phi) = R\lambda \cos\phi; \quad g(\lambda, \phi) = R\phi \ .$$

Cylindrical (Mercator) projection



The Mercator projection is a conformal projection that distorts areas (badly at the poles):

$$f(\lambda, \phi) = R\lambda; \quad g(\lambda, \phi) = R \ln \tan\left(\frac{\pi}{4} + \frac{\phi}{2}\right) .$$

Universal Transverse Mercator (UTM) coordinate system

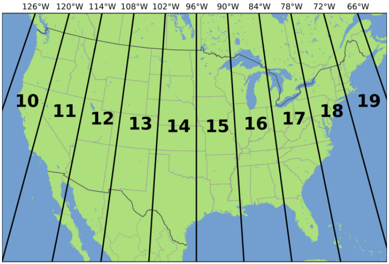
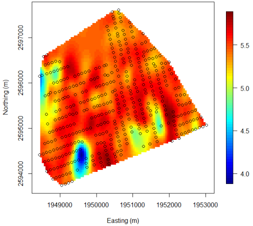


Figure: Simplied figure of UTM grid over USA

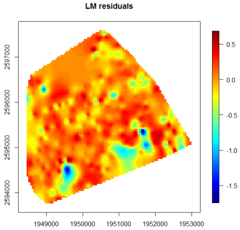
- Popularly used map projection
- Divides the world into 60 north-south zones each of 6 degree longitude
- Within each zone, coordinates are measured north and east in meters

Real data analysis: Bartlett Experimental Forestry data

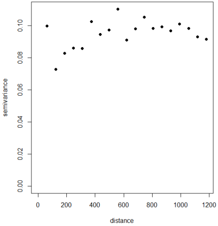
- Forest biomass data for 437 plots in Batlett, NH in 2002
- Predictors: slope, elevation, brightness, greenness and wetness
- Last three predictors obtained from satellite images



(a) log biomass



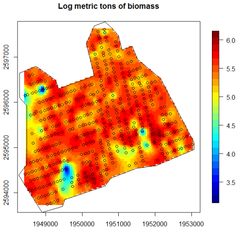
(b) Residuals



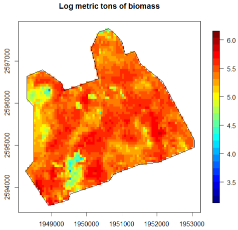
(c) Variogram

Results

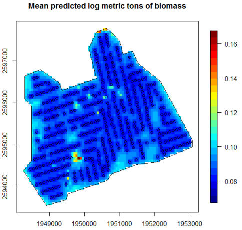
Model	Non-spatial	Spatial
AIC	231	214
BIC	259	250



(d) log biomass



(e) Residuals



(f) Variogram

References

- BCG book chapters 1.2 and 6.3
- Banerjee, S. 2005. *On geodetic distance computations in spatial modeling*. Biometrics, 61: 617–625
- UTM figure: Wikipedia `https://en.wikipedia.org/wiki/Universal_Transverse_Mercator_coordinate_system`