Overview of preferential sampling for spatial data

Brian Reich

9/23/09

SAMSI Working Group on Model Based Geostatistics and Preferential Sampling
Definition

- Diggle, Menezes, and Su: “Preferential sampling arises when the process that determines the data-locations and the process being modelled are stochastically dependent.”

- This dependence is ignored by typical spatial analyses.

- Ignoring this dependence can negatively affect parameter estimation (covariance and mean parameters) and prediction.
Motivating data set #1 - periodontal data

Missing teeth are associated with cumulative periodontal disease

- Maxillary/Lingual
- Maxillary/Buccal
- Mandibular/Lingual
- Mandibular/Buccal

Tooth Number

| Tooth Number | D | MD | MD | MD | MD | MD | MD | MD | MM | DM | DM | DM | DM | DM | DM | DM | DM | DM | DM | DM | DM | DM | DM | D |
|--------------|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----| D |
| 7            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 6            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 5            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 4            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 3            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 2            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 1            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 1            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 2            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 3            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 4            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 5            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 6            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 7            |   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
Motivating data set #2 - scallops data

Fishermen fish where there are many scallops

![Graph showing the relationship between log sampling density and response. The graph includes a 3D scatter plot and a 2D scatter plot with a linear trend line.]
Surely there are many spatial data examples where the presence/absence of an observation is associated with the response.

- Air pollution monitors placed in locations with high pollution.
- Clouds obscuring satellite observations.
- Health of sampled plants or animals.
- Residence of asthma patients.
- Others?
Model for areal data

Reich and Bandyopadhyay, 2009, AOAS

- Let $y_i$ be the response at location $i$ and $z_i$ be the binary indicator of whether $y_i$ is missing.

- Both are modeled in terms of the true value at location $i$, $\mu_i$.

- Gaussian model for the response: $y_i \sim N(\mu_i, \sigma^2)$

- Probit model for missing observations:

  \[ P(z_i = 1) = \Phi(a + b\mu_i) \]

- Spatial model for the truth: $(\mu_1, ..., \mu_n)' \sim \text{CAR}$
Let $s_i$ and $y_i$ be the location and value of the $i^{th}$ observation, respectively.

Both are modeled in terms of the true spatial surface, $\mu(t)$.

Gaussian model for the response: $y_i \sim N(\mu(s_i), \sigma^2)$

Logistic model for the sampling density:

$$p(s_i|\mu) = \frac{\exp(a(s_i) + b\mu(s_i))}{\int \exp(a(t) + b\mu(t)) \, dt}$$

Spatial model for the surfaces: $a, \mu \sim GP$. 
Bayesian extension

- Diggle et al. use an approximate maximum likelihood method.

- Debdeep Pati, David Dunson, and I have been working on a Bayesian version.

- We have shown that an improper prior for $b$ gives a proper posterior under mild assumptions on the prior on $\sigma^2$ and the covariance parameters of the GP prior on $a$ and $\mu$.

- We are close to a proof of a strong consistency of the posterior of $(a, \mu, b)$ assuming infill asymptotics and under standard regularity conditions.
Jointly modeling the response and the sampling distribution presents computational challenges:

- **Probit link**: conjugate and fast, but requires a grid
- **Logit link**: doesn’t require grid, but not conjugate and requires approximation.

A two-stage approach?

- \[ y_i \sim N(\log[\hat{\lambda}(s_i)] + \mu(s_i), \sigma^2), \] where \( \hat{\lambda}(t) \) is an estimate (e.g., kernel density) of the sampling density.

- Our simulations show this often works nearly as well as the fully-Bayesian model. We’re studying the theory.

- Non-linear model: \[ y_i \sim N(g[\hat{\lambda}(s_i)] + \mu(s_i), \sigma^2)? \]

- Spatial propensity score: \[ y_i \sim N(\mu(s_i), \hat{\lambda}(s_i)\sigma^2)? \]
Other comments

How sensitive are these models to simplifying assumptions?

► These models assume normality, linearity, stationarity, etc.

► Connection with the “feedback problem” in the spatial health group? Several people have observed high sensitivity to model misspecification in joint models of spatially-varying exposures and a health response.

When is it necessary to consider preferential sampling?

► Is it asymptotically (infill) irrelevant?

► How does smoothness of the true process affect the bias?

► Can we include the right covariates and then ignore preferential sampling?

► Spatiotemporal data with fixed measurement locations?