An Introduction to Spatial Statistics

Chunfeng Huang
Department of Statistics, Indiana University
Iron Ore (Cressie, 1986)

Raw percent data
Three types of spatial data

- **Geostatistics**
  - Variogram
  - Kriging

- **Lattice or Areal data**
  - Markov Random Field
  - Conditional Autoregressive Model (CAR)

- **Spatial point pattern**
  - Complete spatial randomness (CSR)
  - K function, L function
Geostatistics: Quantify spatial dependency and Assumptions

- Second-order stationarity:
  \[ E(Z(s_i)) = \mu, \quad \text{cov}(Z(s_i), Z(s_j)) = C(s_i - s_j) \]

- Isotropy:
  \[ \text{cov}(Z(s_i), Z(s_j)) = C(||s_i - s_j||) \]

- Intrinsic stationarity:
  \[ E(Z(s_i)) = \mu, \quad \text{var}(Z(s_i) - Z(s_j)) = 2\gamma(s_i - s_j) \]

  \[ \gamma(\cdot): \text{semi-variogram}, \quad 2\gamma(\cdot): \text{variogram}. \]

- Stationary implies intrinsic stationary:
  \[ \gamma(s_i - s_j) = C(0) - C(s_i - s_j) \]

- Intrinsic stationary does not necessarily imply stationary.
vgm(1,"Sph",1.5,nugget=0.2)
Variogram (Semi-variogram)

- Sill, range, nugget
- Conditional negative definite. (Variance may be negative if this is not satisfied.)
- Parametric models
Estimation of variogram

- Empirical variogram (Method-of-moments estimator)

\[ 2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} \left\{ Z(s_i) - Z(s_j) \right\}^2, \]

where \( N(h) \) is the set of pairs with distance \( h \), \( |N(h)| \) is the number of pairs.

- Tolerance regions.

- Fitting parametric models (weighted least squares, MLE, REML)
Empirical variogram

![Empirical variogram graph](image-url)

The graph shows the empirical variogram with distance on the x-axis and semivariance on the y-axis. The data points indicate an increasing semivariance with distance, suggesting spatial autocorrelation in the data.
**Kriging**

Suppose we observe a spatial process $Z(s_1), Z(s_2), \ldots, Z(s_n)$. The best (in terms of minimizing mean squared prediction error) unbiased linear predictor of $Z(s_0)$ is (Cressie, 1993)

$$
\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i),
$$

where

$$
(\lambda_1, \ldots, \lambda_n) = \left(\gamma + \mathbf{1}\left(1 - \mathbf{1}^T \Gamma^{-1} \gamma\right)\right)^T \Gamma^{-1}
$$

$$
\gamma = (\gamma(s_0 - s_1), \ldots, \gamma(s_0 - s_n))^T
$$

$$
\Gamma = \{\gamma(s_i - s_j)\}_{n \times n}
$$

Kriging variance:

$$
\sigma_K^2(s_0) = \lambda^T \Gamma^{-1} \lambda - (\mathbf{1}^T \Gamma^{-1} \mathbf{1} - 1)^2 / (\mathbf{1}^T \Gamma^{-1} \mathbf{1})
$$
\( o = \text{Median Iron Ore \%} \)

\( x = \text{Mean Iron Ore \%} \)
Areal Data or Lattice Data

Simultaneous autoregressive (SAR) model

\[ z(s_1) = b_{12} z(s_2) + \ldots + b_{1n} z(s_n) + \epsilon_1 \]
\[ z(s_2) = b_{21} z(s_1) + \ldots + b_{2n} z(s_n) + \epsilon_2 \]
\[ \vdots \]
\[ z(s_n) = b_{n1} z(s_1) + b_{n2} z(s_2) + \ldots + \epsilon_n \]

Or

\[ z = Bz + \epsilon \]

- Explanatory variables.
- One can show that \( \epsilon \) is not independent of \( z \).
- What if \( z \) is discrete?
Conditional autoregressive (CAR) model

$$z(s_i) \mid \{z(s_j) : j \neq i\} \sim N\left(\sum_{i=1}^{n} c_{ij} z(s_j), \tau_i^2\right)$$

where

$$c_{ij} \tau_j^2 = c_{ji} \tau_i^2, \quad c_{ii} = 0$$

If we further assume the conditional dependency only through the neighborhood of $s_i : N(s_i)$

$$c_{ik} = 0, \ k \notin N(s_i)$$

(Markov random field (MRF))
- Response variable falls in exponential family: auto spatial models.
  - Response variable is normal: auto Gaussian model.
  - Response variable is binary: auto logistic model.
  - Response variable is poisson: auto poisson model.
- SAR can be represented as a CAR model. Not necessarily vice versa, see example in Cressie (1993).
- Markov random field theory is used to guide from the conditional distributions to a valid joint distribution.
Proximity matrix $W$.

Some choices for $W_{ij}$ ($W_{ii} = 0$):

- $W_{ij} = 1$ if $i, j$ share a common boundary.
- $W_{ij}$ is an inverse distance between units.
- $W_{ij} = 1$ if distance between $i$ and $j$ less than some fixed value.
- $W_{ij} = 1$ for $m$ nearest neighbors.
CAR model

\[ z \sim N(X\beta, (I - C)^{-1}M) \]

where

\[ C = \{c_{ij}\}_{n \times n}, \quad M = \tau^2 I \]

and

\[ C = \rho W \]

\[ \rho = 0 \Rightarrow \text{spatial independence} \]

Condition on \( \rho \): \( I - \rho W \) is positive definite.
A spatial point process is said to have the complete spatial randomness (CSR) property if it is a homogeneous poisson point process.

- $A_1, \ldots, A_r$ disjoint, then $N(A_1), \ldots, N(A_r)$ are independent. ($N(A)$: number of events in $A$)
- $N(A) \sim Poisson(\lambda |A|)$, where $|A|$ is the volume of $A$. 
Testing for CSR

- $W$: distance from a randomly chosen event to its nearest event.
- $X$: point to nearest event.
- Through Monte Carlo test.
First order intensity function

\[ \lambda(x) = \lim_{|dx| \to 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\} \]

Second order intensity function

\[ \lambda_2(x, y) = \lim_{|dx|,|dy| \to 0} \left\{ \frac{E[N(dx)N(dy)]}{|dx||dy|} \right\} \]

Stationary: \( \lambda(x) = \lambda \), \( \lambda_2(x, y) = \lambda_2(x - y) \)

Isotropy: \( \lambda_2(x - y) = \lambda_2(||x - y||) \)

Under CSR: \( \lambda(x) = \lambda \), \( \lambda_2(x, y) = \lambda^2 \)
K function

\[ K(t) = \frac{1}{\lambda} E[N_0(t)] \]

where \( N_0(t) \) is the number of further events within distance \( t \) of an arbitrary event.

One can show that

\[ \lambda_2(t) = \frac{\lambda^2 K'(t)}{2\pi t}, \quad \lambda^2 K(t) = 2\pi \int_0^t \lambda_2(y) y dy \]

For the CSR: \( K(t) = \pi t^2 \)

L function:

\[ L(t) = \sqrt{K(t)/\pi} - t \]
References