Spatial Statistics

Ji Meng Loh

New Jersey Institute of Technology

Astrostatistics Summer School
Penn State University
Jun 7, 2013

Outline

Introduction

- Spatial data - data about a phenomenon that includes information about the locations at which the data-points are collected.
- Examples - climate data (rainfall, temperature etc), soil pH in a field/forest, household income
- Usually, data collected from nearby locations are more similar - that is, the data is dependent or correlated.
- More specific than applications with general dependent data. The dependence is due to spatial locations - usual assumption, dependence decreases with distance apart.

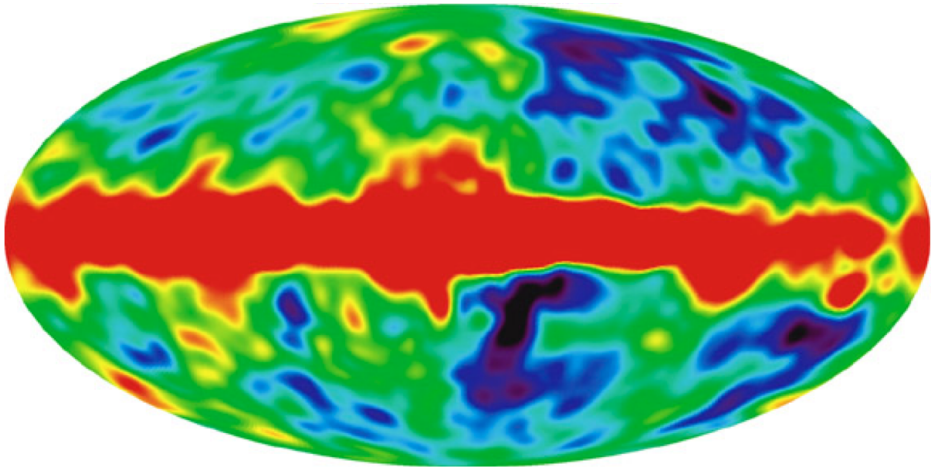
Accounting for dependence

- Key feature in spatial data is the dependence/correlation in the observations
- Need to account for the dependence in statistical models.
- If assume independence, we are assuming we have more information than we really do - underestimate the uncertainty
- Accounting for dependence can help with better estimates and predictions.
- Often reasonable to assume that the dependences decreases with distance apart

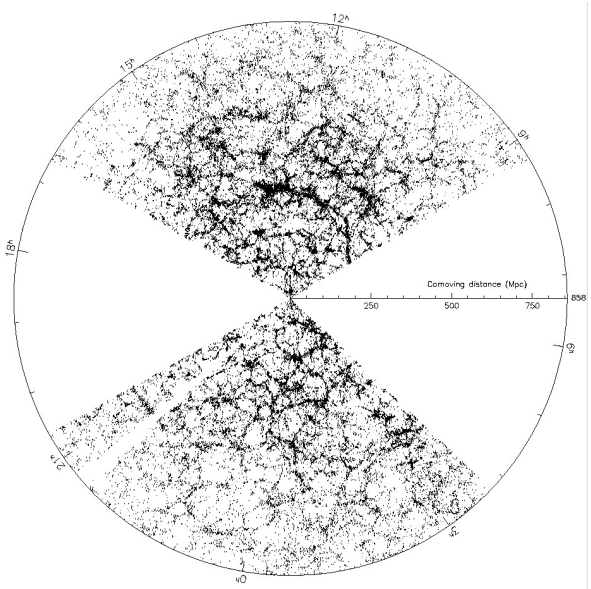
Types of spatial data

- Geostatistical - spatial process in continuous space, with observations at specific locations, e.g. ozone measured at a number of monitoring stations; mining application
- Spatial point processes - locations of objects in space, e.g. trees in a forest, galaxies in space
- Lattice data - data (usually counts) observed on a regular or irregular grid, e.g. census data, counts by zip-code, brain maps

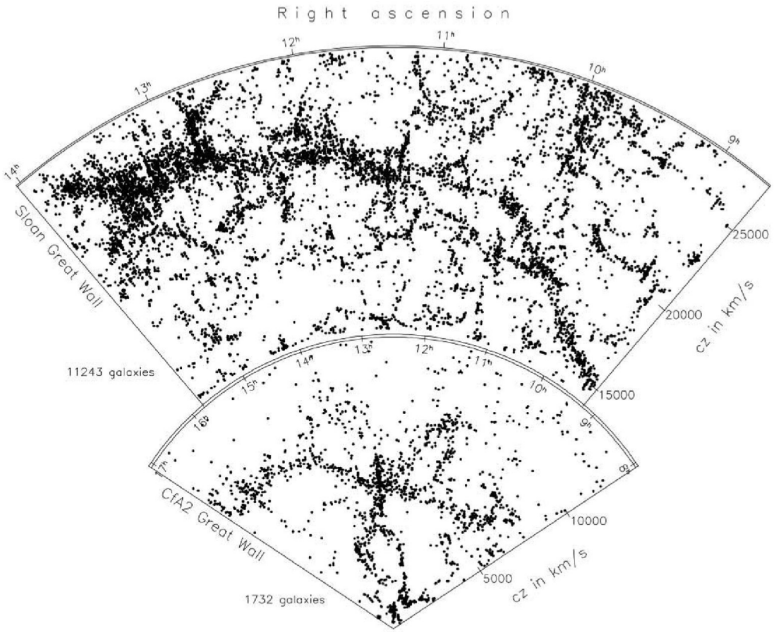
CMB



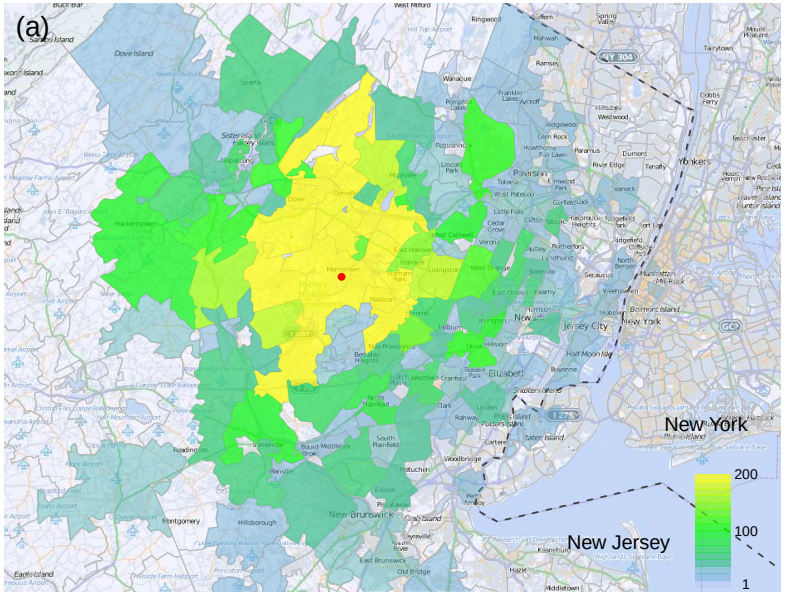
SDSS



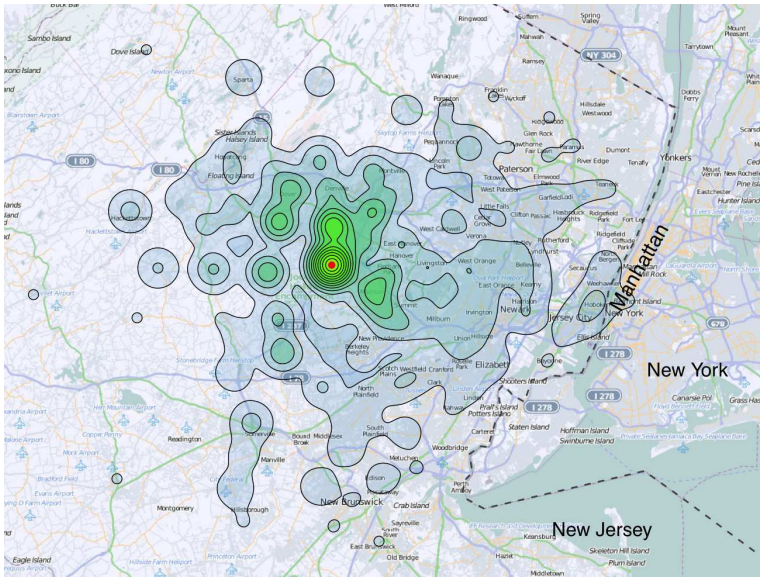
SDSS and CfA2



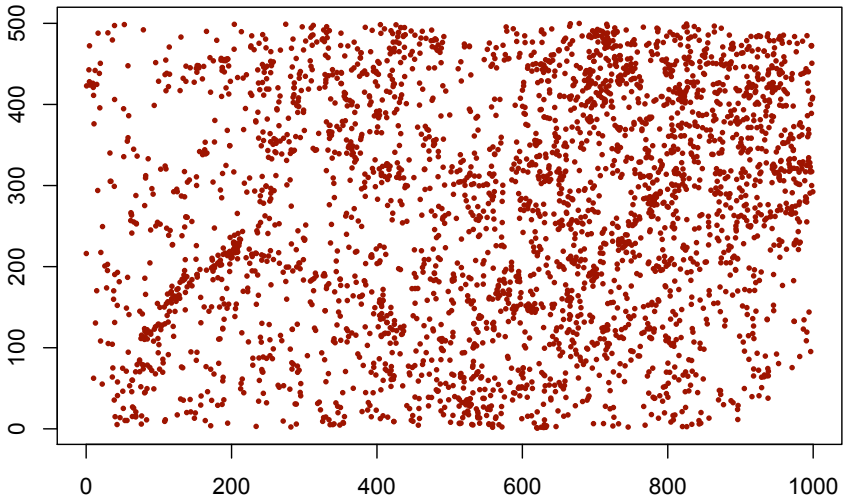
Where do people who work in Morristown live?



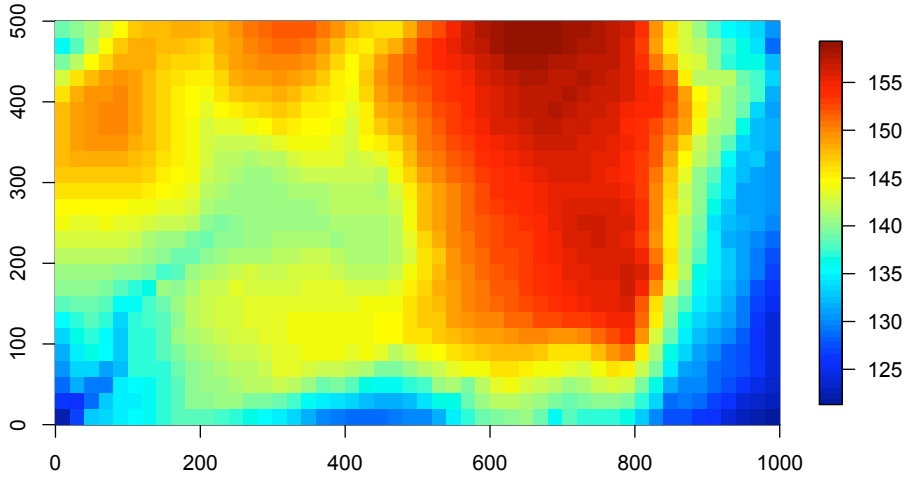
Morristown laborshed



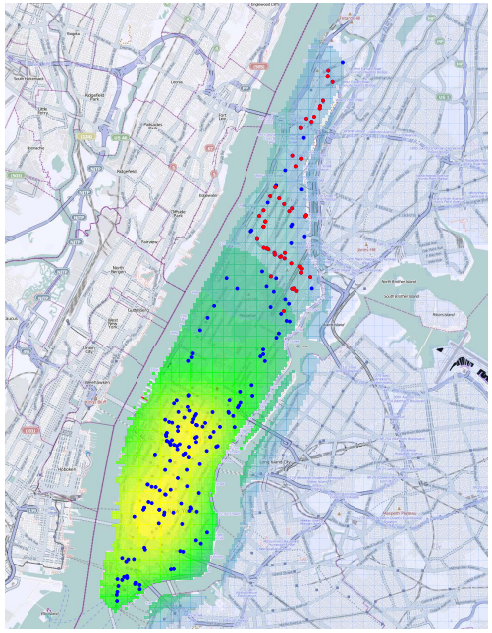
Locations of a tree species in plot of land



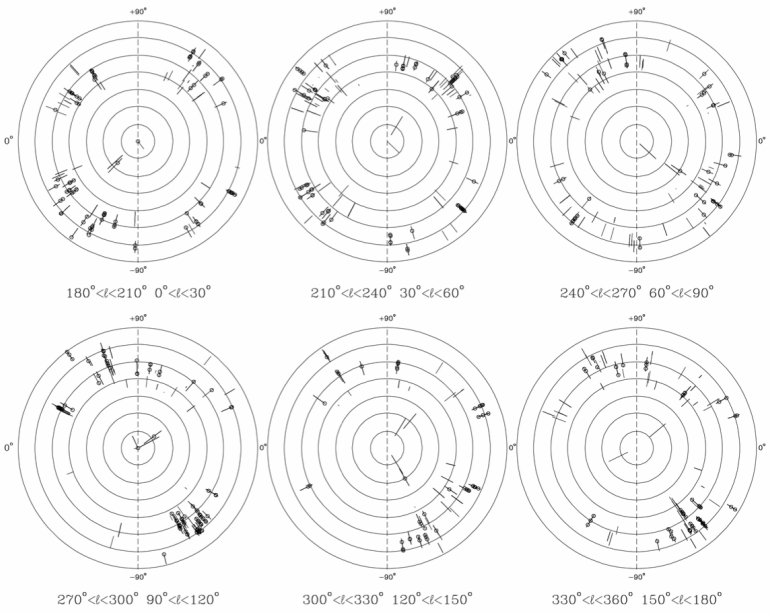
Measurements of elevation in the same plot of land



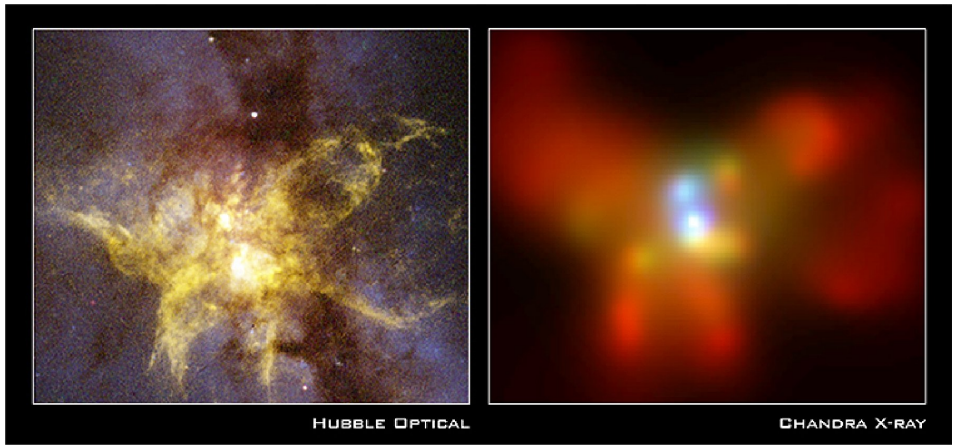
Fastfood restaurants in NYC



C IV absorption systems



Images



Outline

Geostatistics

- Random fields - basic properties; Gaussian Random Fields
- Mean function and error covariance structure
- Some models for covariance function
- Estimation - variogram, REML, ML
- Kriging

Random fields

Spatial process Z observed at locations s in continuous space, $s \in D \subset \mathbb{R}^d$.

- $Z(s) = \mu(s) + \epsilon(s)$
- $\mu(s) = \mathsf{E}[Z(s)]$ is the mean
- $\epsilon(s)$ is the spatial error - model dependence through ϵ.
- Define covariance function $C(s,t) = \mathsf{Cov}(Z(s), Z(t))$
- $C(s,t)$ has to be positive definite: $\forall n, s_1, \; dots, s_n \in D \subset \mathbb{R}^d$ and $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} c_i c_j C(s_i, s_j) > 0$$

Stationarity, isotropy and Gaussianity

Usually make some assumptions to simplify the model:

- Stationarity - properties of process are the same wherever you are: $C(s, s+h) = C(h)$
- Isotropy - properties of process are the same in any direction: $C(s, s+h) = C(|h|)$
- Gaussianity - $\forall n, s_1, \ldots, s_n \in D \subset \mathbb{R}^d$, $Z(s_1), \ldots Z(s_n)$ is multivariate Gaussian distribution, i.e.

$$\mathbf{Z} \sim N(\mu, \Sigma)$$

- if covariates $\mathbf{X}$ are measured in D, can set $\mu = \mathbf{X}\beta$
- Also, specify a parametric form for Σ that ensures that Σ is positive definite.

Models for the covariance function

- Exponential:

$$C(h) = \sigma^2 \exp\{-\theta h\}, \quad \theta > 0$$

- Gaussian:

$$C(h) = \sigma^2 \exp\{-\theta h^2\}, \quad \theta > 0$$

- Spherical:

$$C(h) = \sigma^2(1 - 3h/2\alpha + (h/\alpha)^2/2), \quad h \leq \alpha$$

- Matérn:

$$C(h) = 2\sigma^2(\theta h/2)^\nu K_\nu(\theta h)/\Gamma(\nu), \quad \nu > 0, \theta > 0$$

K is a Bessel function

Variogram

The variogram is a measure of the spatial dependence:

$$\begin{aligned} 2\gamma(s_i - s_j) &= \text{Var}[Z(s_i) - Z(s_j)] \\ &= \text{E}[(Z(s_i) - Z(s_j))^2] \quad \text{if mean is constant} \end{aligned}$$

$\gamma(s_i - s_j) \equiv \gamma(|s_i - s_j|)$ under stationarity and isotropy
Simple estimator:

$$2\gamma(\hat{h}) = \sum_{N(h)} [Z(s_i) - Z(s_j)]^2 / |N(h)|,$$

$N(h)$ is set of all pairs (s_i, s_j) with $|s_i - s_j| \in (h - \Delta h, h + \Delta h)$.

Inference

- Match estimated variogram with model variogram (exponential, Matérn etc) by (weighted) least squares
- Likelihood methods - Gaussian likelihood $\mathcal{L}(\theta, \beta; \mathbf{Z})$
  - regular maximum likelihood
  - Restricted maximum likelihood (REML) using likelihood based on contrasts
  - Composite likelihood - full likelihood of groups of observations, independence between groups
- Bayesian methods

With estimates of the parameters, can then use the model to make predictions at other locations with no observations (kriging)

Computing

- R packages - geoR, fields, RandomFields, spBayes
- WinBugs, JAGS for MCMC
- Matrix computations limit the data size that standard methods and R packages can handle

Example - modeling magnitudes of Type Ia Supernovae

- "Standardizing Type Ia Supernova Absolute Magnitudes Using Gaussian Process Data Regression" - A. Kim et. al (http://arxiv.org/abs/1302.2925)
- Given time measurements of supernova photometry in 4 bands, want to estimate/predict the absolute magnitude
- Borrow strength from other supernova data by using a Gaussian process model
- Fill in the gaps in the multi-band light curves
- Reduce dimension using principal components analysis (PCA)
- Relate absolute magnitude to PCA coefficients; can then predict abs. magnitude given light-curve shape and colors.

Gaussian process model

- At epoch t, filter λ, photometric magnitude $m_{(t,\lambda)}$ is given by

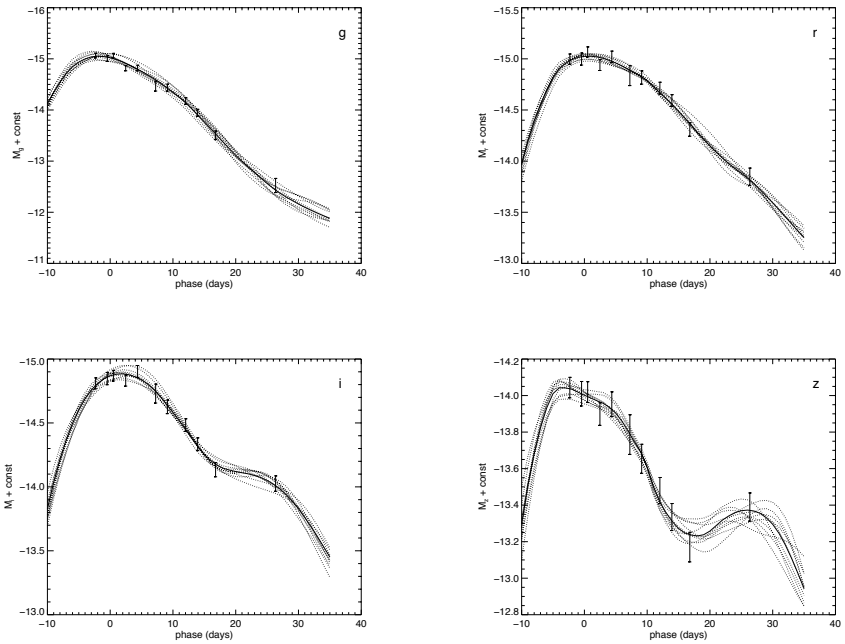
$$m_{(t,\lambda)} \sim \text{GP}\left(\bar{m}(t,\lambda;m_0), k_m(t,\lambda,t',\lambda',l_{k_m},\sigma_{k_m})\right)$$

- $\bar{m}$ based on templates,

$$k_m(t,\lambda,t',\lambda',\theta) = \sigma^2_{k_m}(\lambda)\delta_{\lambda\lambda'}\exp\left[-\left(\frac{t-t'}{l_{k_m}}\right)^2\right]$$

- Note that the k_m does not correlate across bands
- The parameters l_{k_m}, σ_{k_m} and other parameters are estimated from data

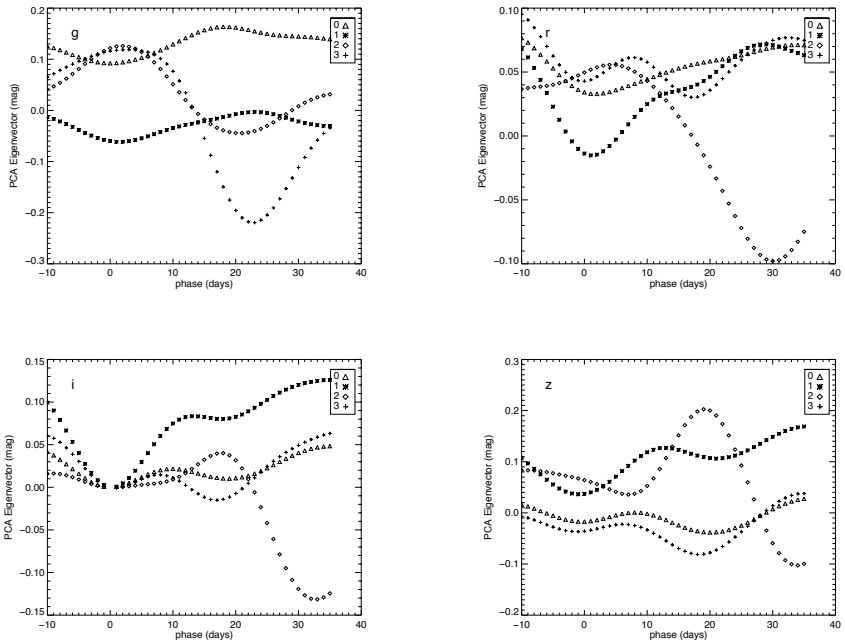
Predicted mean plus ten interpolated light curves



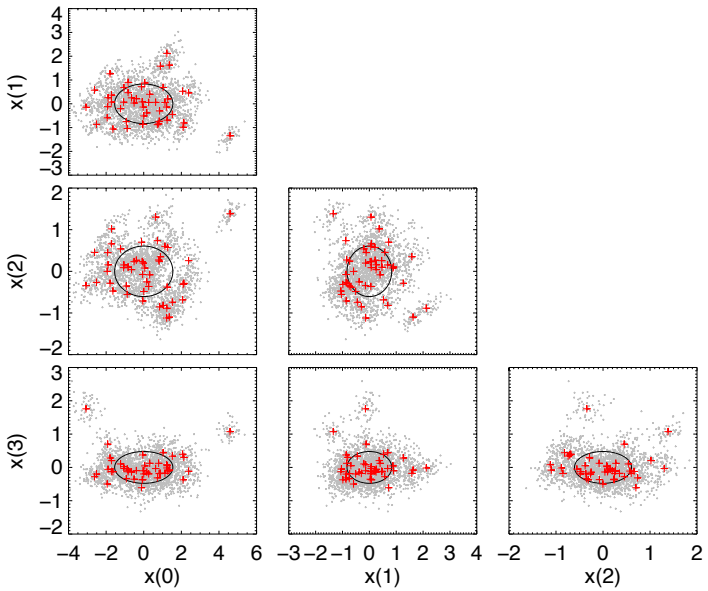
Principal components

- Express light curves in terms of principal components that account for 95% of the variance
- Have 4 eigenvectors and for each light curve, a set of 4 eigenvalues for each band
- Also have values of absolute magnitude for these data

First 4 principal components in the 4 bands



Distribution of PCA coefficients



Regression models

- With $x(j)$ representing the j-th eigenvalue for light curve x, can then build a model between absolute magnitude of supernova generating x and $x(j)$
- Linear model:

$$M_\lambda(x;\mathbf{p}) = \bar{M}_0 + \sum_{j=1}^{N_p} p_j x(j)$$

- Gaussian model:

$$M_\lambda \sim \text{GP}\left(\bar{M}, k_M\right),$$

$\bar{M}$ based on linear model, k_M similar to previous GP model, but depending on $x(j)$'s.

Outline

Lattice data - Markov random fields

- Random process Z observed on locations s on a lattice.
- Examples - disease counts by zip code; intensity at pixels in an image.
- Usual to specify the dependence conditionally, i.e. we model the conditional distribution of $Z(s_i)$ given $Z(s_j), j \neq i$.
- Simplify the above by assuming a Markov property - $Z(s_i)|Z(s_j), j \neq i$ becomes $Z(s_i)|Z(s_j), j \sim i$, where $j \sim i$ means s_j is a neighbor of s_i, i.e. process at a location s is independent of other locations *given* process at neighboring locations s'.

Gaussian Markov random fields

- Conditional distributions are Gaussian, resulting distribution is multivariate Gaussian
- $Z_i|Z_j, j \neq i, \tau \sim N(n_i^{-1} \sum_{j \sim i} Z_j, (n_i \tau)^{-1})$
- With Θ the parameters of the precision matrix $Q(\Theta)$:

$$\mathbf{Z} \sim N(\mathbf{X}\beta, Q^{-1}(\Theta))$$

- Estimate Θ, β using MLE or Bayesian methods
- Because the matrices tend to be sparse (due to the Markov property), computation is less challenging than with Gaussian processes.

Software

- GeoDa R package by Luc Anselin
- spdep R package by Roger Bivand
- INLA by Havard Rue - includes stochastic partial differential equation (SPDE) approach to approximate Gaussian Processes using GMRFs
- GeoBUGS which is part of WINBUGS

Outline

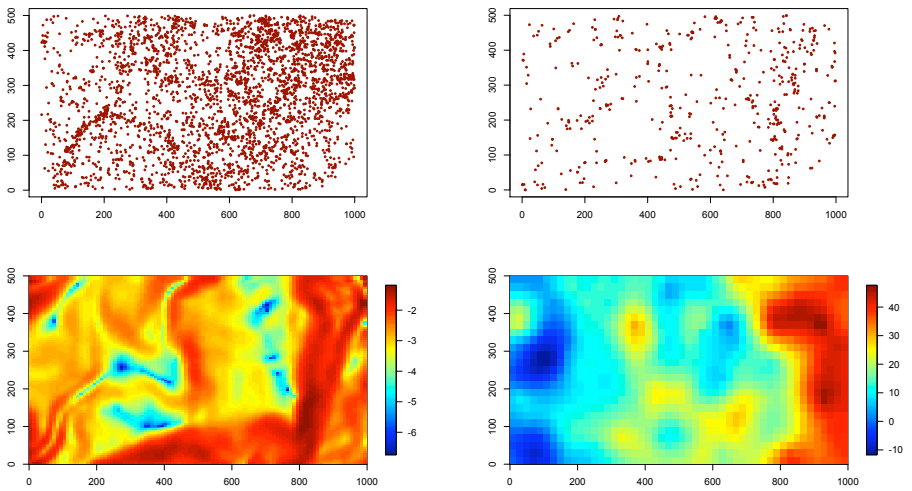
Point patterns

- Locations of objects/points in space, e.g. trees in a forest, locations of galaxies
- Think of a random mechanism underlying the process that can generate many point patterns. But often our data is only one realization, i.e. no replication
- Observation region/window is the region where the objects can potentially be found. Lack of points within the observation region provides information about the point process.
- Can also think of it as a continuous random process X taking values 0 or 1, observed at all points in the observation region.

Some questions

- How dense are the points? Number per unit area/volume: Number density (also called intensity)
- How clumpy or spread out (regular) are the points, i.e. do they interact with one another? Correlation between where points occur - 2PCF, K function etc
- Homogeneous (stationary), i.e. the statistical characteristics stay the same in different locations, or inhomogeneous (non-stationary)?
- Are there other measurements made of the points? For example, width/height of the trees, magnitude of the galaxies.
- Are there measurements of other variables within the observation region? For example, pH of soil

Locations of two tree species; measures of Elevation and Nitrogen



Some technical details

- A spatial point process Φ is a stochastic process giving realizations $(x_1, \ldots, x_n)$ in a region $W \subset \mathbb{R}^d$
- Note that n is random
- For a region $A \subset S$, define $N(A) = \#$points in A.
- $N(A)$ is random. Define $\mathsf{E}(N(A))$ to be $\Lambda(A)$ for any $A \subset S$.
- If $\lambda(\cdot)$ exists such that $\Lambda(A) = \int_A \lambda(s)\, ds$, λ is called the intensity function.

$$\lambda(s) = \lim_{|ds| \to 0} \frac{\mathsf{E}(N(ds))}{|ds|}$$

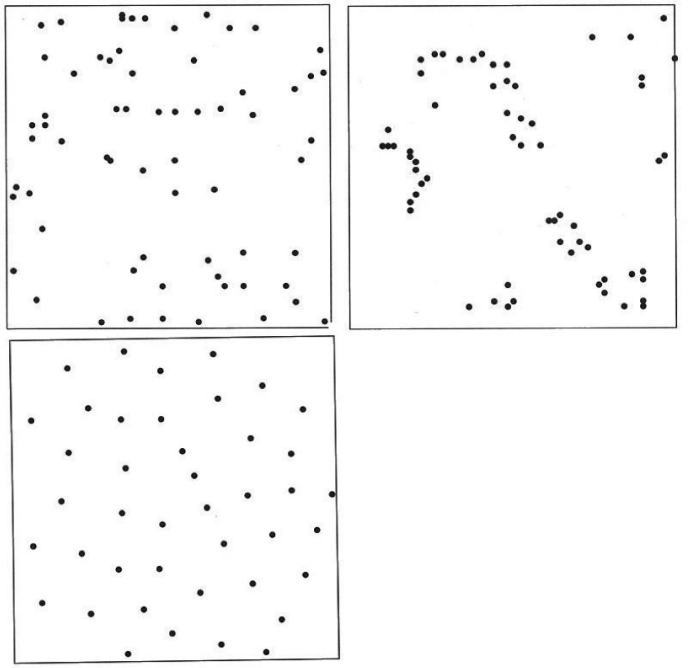
More technical details - stationarity and isotropy

- Stationarity: process $\Phi = \{x_n\}$ and translated process $\Phi_x = \{x_n + x\}$ have the same distribution for all x
- "different regions of the observation region yield similar configurations of points"
- Difficult to prove - stationary process can look non-stationary within a bounded window and vice versa
- Isotropy: process $\Phi = \{x_n\}$ and rotated process $R_\alpha \Phi = \{R_\alpha x_n\}$ have the same distribution for all α

Some stationary point models

1. Poisson with constant intensity: independence between points, just specified by the intensity
2. Neyman-Scott: parent-offspring model; locations of parent points are Poisson spatial point process, Poisson number of offspring are distributed say, uniform on a disc centered about parents
3. Inhibition processes, soft-core and hard-core processes
4. Gibbs or Markov processes: specified by interaction function between pairs of points (or groups of n points, $n = 3, ...$); Strauss process is an example

Simulated realizations



A bit more on the Poisson process

- For $B \subset W$, $N(B) \sim$ Poisson$(\Lambda(B))$
- Given number of points $N(W)$, the point locations are randomly and independently distributed in W according to pdf proportion to $\lambda(s)$.
- For two disjoint regions, B_1, B_2, $N(B_1)$ and $N(B_2)$ are independent.
- Cox processes - random intensity function; given the intensity function, the process is inhomogeneous Poisson.

Log Gaussian Cox Process

- random intensity $\Lambda(s) = \log Z(s)$ where $Z(s)$ is a Gaussian process
- Suppose Z is stationary and has mean μ, variance σ^2 and correlation function $\rho(t)$. Then,

$$\begin{aligned} \lambda &= \exp(\mu + 0.5\sigma^2) \\ \text{Cov}(\Lambda(s_1), \Lambda(s_2)) &= \exp(\sigma^2 \rho(t)) \end{aligned}$$

- Used by Coles and Jones (1991) - MNRAS 248 "A lognormal model for the cosmological mass distribution"

Inhibition processes

- Matérn hard core process: Poisson process with intensity ρ, thinned by deleting all pairs of events less than δ apart
- Points of Poisson process marked independently with times of birth (say uniform (0,1)). An event is removed if it lies within distance δ of an older event
- Simple sequential inhibition process: put a sequence of events in W, given $\{x_j : j = 1, \ldots, i-i\}$, x_i is uniformly distributed on $W \cap \{y : |y - x_j| \geq h, j = 1, \ldots, i-1\}$

Gibbs/Markov process

- Models for regular patterns that are more flexible than inhibition processes
- For a point configuration χ, let $f(\chi)$ represent how much more likely the configuration is for that process, compared with the Poisson process with intensity 1.
- Turns out that $f(\chi)$ can always be factorized into

$$f(\chi) = \alpha \prod_{i=1}^{n} g_i(x_i) \prod_{j>i} g_{ij}(x_i, x_j) \ldots$$

- Pairwise interaction processes:

$$f(\chi) \propto \prod_{\xi \in \chi} \lambda(\xi) \prod_{(\xi,\eta)} \phi(\xi, \eta)$$

f is repulsive if $\phi < 1$ homogeneous if λ is constant and $\phi(\xi, \eta) = \phi(\xi - \eta)$.

Strauss process

- $\phi(r) = \gamma^{1\{r \leq R\}}, \quad \gamma \in [0, 1], R > 0$; R is the range of interaction
- $f(x) \propto \beta^{n(x)} \gamma^{s(x)}$, $n(x)$ is number of points, $s(x)$ is number of R-close pairs
- $\gamma = 1$ gives the Poisson processs; $\gamma < 1$ has repulsion between R-close pairs; $\gamma = 0$ gives a hard core process with core distance R.

Given a spatial point dataset, what can we do?

- Test for Complete Spatial Randomness (CSR) or Poissonity: quadrant count method, and various distance methods (distribution of event-event distances, nearest neighbor distances, point-event distances)
- Estimate the constant intensity: $\hat{\lambda} = N(W)/|W|$
- Estimate the second-order properties: K function, pair correlation function, two-point correlation function
- Fit a model
- Model the inhomogeneous intensity as a function of measured covariates

Second-order properties

- Second-order intensity function of a spatial point process:

$$\lambda^{(2)}(s_1, s_2) = \lim_{|ds_1|\to 0} \lim_{|ds_2|\to 0} \frac{\mathsf{E}[N(ds_1)N(ds_2)]}{|ds_1||ds_2|}$$

- Under stationarity, $\lambda(2)(s_1, s_2) = \lambda^{(2)}(s_2 - s_1)$.
- Under stationarity and isotropy, $\lambda^{(2)}(s_1, s_2) = \lambda^{(2)}(|s_2 - s_1|)$.
- $\lambda^{(2)}$ is hard to interpret; second-order product density $\rho^{(2)}$ is a bit easier: $\rho^{(2)}(s_1, s_2)ds_1\, ds_2$ is the probability of finding a point each in the infinitesimal volumes at s_1 and s_2.

Second-order properties (cont)

- Pair correlation function: $g(r) = \rho^{(2)}(r)/\lambda^2$
- Ripley's K function: $\lambda K(r)$ is the expected number of points within distance r of a randomly selected point.
- In $\mathbb{R}^2$, $K(r) = \pi r^2$ for the homogeneous Poisson process
- Some formulas (for $\mathbb{R}^3$):

$$g(r) = \frac{1}{4\pi r^2}\frac{dK(r)}{dr} \qquad \rho^{(2)}(r) = \frac{\lambda^2}{4\pi r^2}\frac{dK(r)}{dr}$$

$$L(r) = \left(\frac{K(r)}{\pi}\right)^{1/2}$$

Estimating the K function

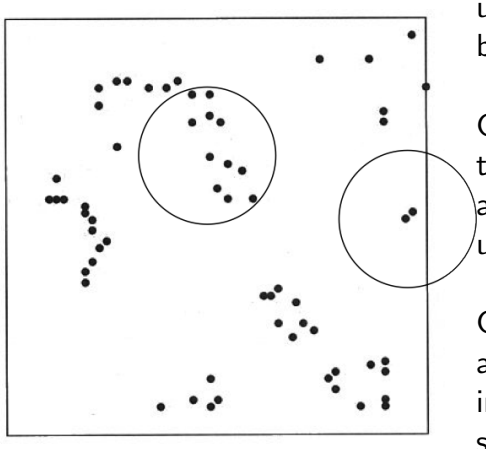
- Usually, estimate $\lambda^2|W|K(r)$ then divide by estimate of $\lambda^2|W|$, i.e. $N^2/|W|$ or $N(N-1)/|W|$.
- Naive estimator of $\lambda^2|W|K(r)$:

$$\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} 1_{(0,r]}(|x_i - x_j|)$$

  Does not take into account the boundary of the observation region, hence is an under-estimate.

- Can restrict points i to be far enough from the boundary, but discards some information, especially for large r.

Edge effects



Neighborhood around points near the boundary is partially unobserved and may create bias in estimates

Can use a buffer zone to avoid the boundary, or use correction adjustments to account for unobserved events

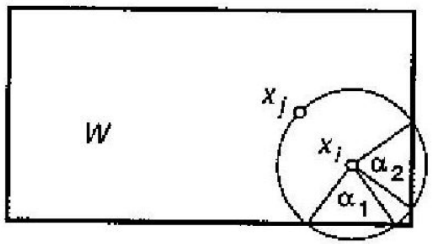
Correction adjustments allow all the data to be used, eliminate bias, but increases the standard errors

Estimating the K function with edge effects

- Essentially, give each pair of points a weight: if a point pair is near a boundary, give it a higher weight.
- Ripley's estimator:

$$\sum\sum 1_{(0,r]}(|x_i - x_j|)b_{ij},$$

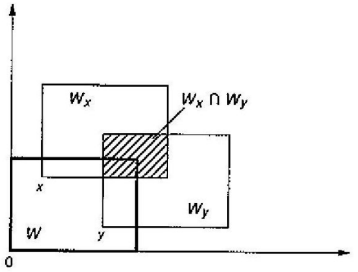
where b_{ij} represents the proportion of the circle of radius $|x_i - x_j|$ centered at x_i that is in $|W|$



Estimating K (cont)

- Translation estimator:

$$\sum\sum 1_{(0,r]}(|x_i - x_j|)/|W_{x_i} \cap W_{x_j}|$$



- Rotation (Ohser's) estimator: use an averaged version of $|W_{x_i} \cap W_{x_j}|$

Two-point correlation function

- $\xi(r) = g(r) - 1$
- "over-density" - excess probability of finding 2 points separated by distance r, compared to the Poisson
- Astronomers tend to use the 2PCF, estimating it using a numerical method to account for edge effects

Astronomers' estimates of the 2PCF

- D: dataset, R: random set

$$\begin{aligned} DD(r) &= \sum_{x\in D}\sum_{y\in D} 1\{|x-y| \in (r-\Delta r, r+\Delta r)\} \\ DR(r) &= \sum_{x\in D}\sum_{y\in R} 1\{|x-y| \in (r-\Delta r, r+\Delta r)\} \end{aligned}$$

- Some estimators:

$$\hat{\xi}_N = \frac{DD}{RR} - 1; \quad \hat{\xi}_{DP} = \frac{DD}{DR} - 1; \quad \hat{\xi}_{He} = \frac{DD - DR}{RR}$$

$$\hat{\xi}_{Ham} = \frac{DD \times RR}{DR^2} - 1 \quad \hat{\xi}_{LS} = \frac{DD - 2DR + RR}{RR}$$

Estimating errors

- Poisson errors (Ripley 1988; Landy and Szalay 1993)
- Parametric bootstrap - simulate from a model of the process (Einsenstein et al. 2005) and estimate quantity from simulated data
- Non-parametric bootstrap - block bootstrap, marked point bootstrap
- Block bootstrap - randomly place blocks in the observation region to copy the point pattern; join multiple blocks together to get a "new" sample; estimate quantity from the resamples

Marked point bootstrap

- Estimate is of the form $\sum_i \sum_{j \neq i} 1_{(0,r]}(|x_i - x_j|) w_{ij} \equiv \sum_i m_i$
- Assign m_i to each point x_i as marks
- Copy points using blocks, but instead of making a new point pattern and re-computing estimate from scratch, just add up the resampled marks m_i^*.

Fitting a model

Likelihood methods tend to be difficult except for the Poisson or Markov models. A simple way is to fit using minimum contrast:

With a point model with parameter(s) θ and theoretical $K(r;\theta)$ function, can estimate θ by finding the value that minimizes

$$D(\theta) = \int_0^{r_0} w(t)[\hat{K}(r)^c - K(r;\theta)^c]^2 \, dr$$

Find standard errors by simulating realizations from the fitted model and repeating the above procedure for the simulated realizations.

Estimating the intensity function

- Using a kernel with bandwidth h:

$$\hat{\lambda}_h(x) = \sum_{i=1}^{n} k_h(x - x_i)$$

- Examples of k_h:

$$k_h(z) = 1_{b(0,h)}(z)/\pi h^2$$

$$k_h(z) = 8e_h(||z||)/3\pi h \quad e_h(t) = \begin{cases} \frac{3}{4h}\left(1 - \frac{t^2}{h^2}\right) & |t| < h \\ 0 & \text{otherwise} \end{cases}$$

- Optimal h can be chosen using cross-validation methods.

Estimating the intensity function (cont)

- for an inhomogeneous Poisson point process X with intensity function $\lambda(s)$, the log-likelihood is given by

$$l(\lambda; X) = \sum_{i=1}^{N} \log \lambda(x_i) - \int_W \lambda(s)\, ds$$

- Intensity function can be expressed in terms of covariates and/or coordinates (s_1, s_2):

$$\begin{aligned} \log \lambda(s) &= \sum_{j=1}^{p} \beta_j z_j(s) \quad \text{or something like} \\ \log \lambda(s) &= \alpha + \beta_1 s_1 + \beta_2 s_2 + \gamma_1 s_1^2 + \gamma_2 s_2^2 \end{aligned}$$

The z's have to be observed at more locations than just the event locations.

Poisson likelihood as estimating equations

- Even for non-Poisson processes, can use the Poisson likelihood to model the intensity function. Schoenberg (2005) provided conditions for the estimates of θ to be consistent.
- Can fit using the *ppm* function in the *spatstat* R package
- Waagepetersen (2007) introduced an inhomogeneous Neyman-Scott process (INSP) by randomly thinning (removing) points from a homogeneous NS process. Parameters for the intensity are estimated from the Poisson estimating equations, parameters for the Neyman-Scott process estimated using minimum contrast for the K function.

An application

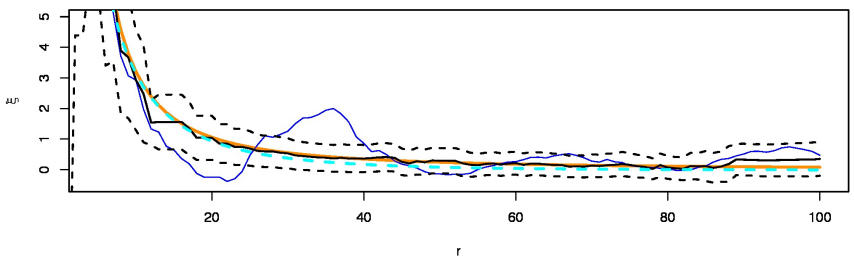
Estimation of K function or 2PCF requires a bin size. Using optimal bin sizes can reduce standard errors. May also expect optimal bin size to vary with distance separation r.

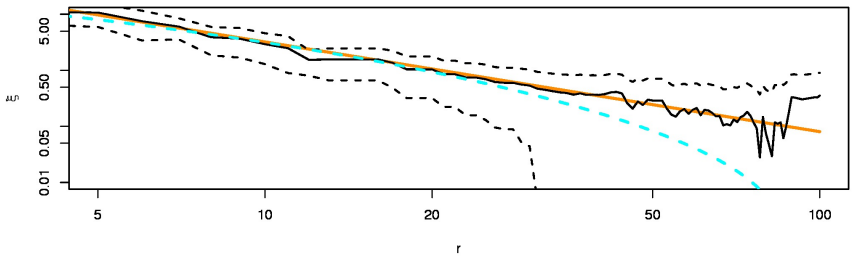
A bootstrap bandwidth selection method:

- $\hat{\xi}_p(r)$ is a pilot estimate of ξ; $\hat{\xi}_h(r)$ is estimate using bin size(s) h
- $\hat{\xi}_b^*(r), b = 1, \ldots, B$ are bootstrap estimates of ξ
- Find h that minimizes

$$\sum_r [\hat{\xi}_p(r) - \hat{\xi}_h(r)]^2 + \text{Var}(\hat{\xi}^*)$$

An application (cont)



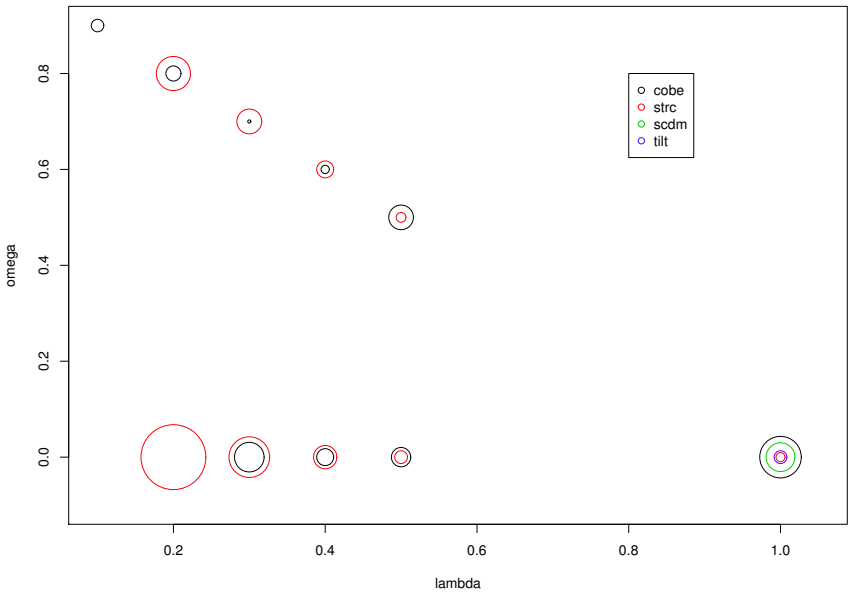


An application (cont)

- Use optimal bin sizes to estimate 2PCF for, say, SDSS LRGs
- Do the same for mock catalogs from various cosmologies
- Compute discrepancy D:

$$D = \sum_r \left(\frac{\hat{\xi}(r) - \hat{\xi}(r;\theta)}{\hat{\sigma}} \right)^2$$

Applied to 2dF GRS and 2dF mock catalogs



References

- Statistics for Spatial Data - Cressie (comprehensive reference)
- Spatial Statistics - Ripley
- Introduction to Geostatistics - Kitanidis (readable intro, based on hydrology)
- Hierarchical Modeling and Analysis for Spatial Data - Banerjee, Carlin and Gelfand (Bayesian)
- Gaussian Markov Random Fields - Rue and Held
- Stochastic Geometry and its Applications - Stoyan, Kendall and Mecke
- Modern Statistics for Spatial Point Processes - Moller and Waagepeterson
- Statistics of the Galaxy Distribution - Martinez and Sarr