

Problems in Spatial Data Analysis

Haining

Chapter 2

SIE 512 Lecture 7

September 2009

Problems in spatial data analysis

Properties of spatial data can create problems for statistical analysis.

- Conceptual models and inference frameworks
- Effects of properties of spatial surfaces on analysis
- Special problems of point and area data

Conceptual Models and Inference Frameworks

Classical inference model - data are the result of well defined experiments that can be repeated numerous times

Statistical inference is inference about a population from a random sample drawn from it or, more generally, about a random process from its observed behavior during a finite period of time.

How should we think about the concept of random samples and inference in a spatial context?

Inference Objective

Make statements regarding properties of the underlying spatial process responsible for the observed data

Conceptual Models and Inference Frameworks

- Design-based inference – the spatial pattern is assumed fixed and chance enters in the choice of sampling pattern.
- Model-based or super-population inference - the sampling pattern may be fixed but the spatial population is assumed to be the outcome of a repeatable chance process.

Difference lies in the assumption of the existence of a universe of possible random surfaces.

Model-based (super-population) inference

In the model based approach we typically have a single realization (experiment) of a random process

From this one realization the expectation is we have sufficient information to identify properties of the underlying surface.

Assuming a form of stationarity, the single realization provides enough information to identify properties of the underlying process

- Strict stationarity requires that both the mean and variance – covariance properties be stationary
- A weaker form of stationarity allows the mean to vary as a function of location but subject to dependence on a small number of parameters
- Regionalized variable theory requires only surface increments to be stationary

Model-based (super-population) inference

Is the hypothetical universe of realizations (super-population) from which realizations are drawn supportable?

We may imagine other “replicates” arising at other times in the same location or in other areas

In the model-based approach the population is considered a random sample of a super-population described as a random function, i.e. as a set of random variables $\{Z_{(s1)}, Z_{(sn)}\}$

The concept of a confidence interval is not built around a sample mean.

The “confidence interval” depends on specification of a model (Z) used to describe how the variable Z is distributed.

Design-based inference

Design-based inference is based on hypothetical repetitions of the sampling process with the spatial population remaining fixed.

Data represent sample observations from a given surface

Spatial surface is continuously varying but fixed (not a realization)

All variation is associated with variation in this surface and random errors

Random components are sampling error and measurement error

Conceptual Models and Inference Frameworks

Either approach assumes data arise from a form of controlled experiment

This in turn assumes axiom of correct specification

- The set of explanatory variables thought to determine the response variable must be unique, complete, small in number, and observable.
- Other determinants of the response variable must have a probability distribution with at most a few unknown parameters
- All unknown parameters must be constant

Need to consider the extent to which the data situation meets these assumptions

Conceptual Models and Inference Frameworks

In non- experimental case of data analysis - there is no experiment that defines the model (e.g. regression model)

Conclusions and inferences can differ between models arrived at under different circumstances

- Clarify purpose of analysis and judgmental basis (Leamer)

Exploratory analysis approaches to non-experimental explanatory inference:

bootstrapping (taking random samples from observations to construct confidence intervals)

randomization – permutations of spatial arrangement

Problem with randomization - all permutations considered equally likely and the observed spatial structure is ignored

Modeling Spatial Variation

Need models that describe spatial variation

What are important characteristics of spatial variation?

Contrast of space and time:

Similarities

- ordered
- not independent
- have different scales of variation
- can be stationary or non-stationary

Differences:

- spatial dependencies typically 2 and 3 dimensional
- ordering - non directional, ambiguous since there are multiple paths
- boundaries are larger, effect more sites
- boundary conditions can have very complex form
- assumptions about the boundary more important - arbitrary, natural, intermediate
- discontinuities are more complex

Statistical modeling of spatial data

Some distinctive problems for point and area referenced data

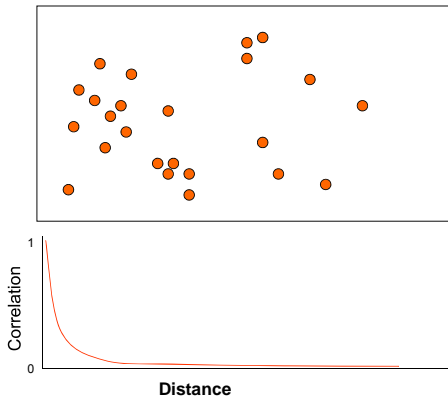
Dependency in spatial data

dependency is similar to duplicating observations already in a data file

dependence represents a loss of information compared to a set of independent observations

data dependency effects the mean, variance about the mean and so effects confidence intervals and significance testing

Dependency in spatial data



Dependency in spatial data

Consider estimation of a constant spatial mean from a set of observations from a normal distribution with mean μ and variance σ^2

Maximum likelihood estimator of the mean

$$\hat{\mu} = (1/n) \sum_{i=1 \dots n} y_i = (\mathbf{1}^T \mathbf{1})^{-1} (\mathbf{1}^T \mathbf{y})$$

Now assume vector \mathbf{y} is a sample from a multivariate normal distribution with constant mean and dispersion matrix $\sigma^2 \mathbf{V}$

Maximum likelihood estimator is now

$$\hat{\mu} = (\mathbf{1}^T \mathbf{V}^{-1} \mathbf{1})^{-1} (\mathbf{1}^T \mathbf{V}^{-1} \mathbf{y})$$

Dependency in spatial data

let $\sigma^2 \mathbf{V} = \sigma^2 (1 - \tau \mathbf{W})^{-1}$ $\mathbf{W} = \{w_{ij}\}$ where w_{ij} is between 0 and 1 and depends on whether i and j are neighbors

$$\hat{\mu} = (\mathbf{1}^T \mathbf{V}^{-1} \mathbf{1})^{-1} (\mathbf{1}^T \mathbf{V}^{-1} \mathbf{y})$$

$$(\mathbf{1}^T \mathbf{V}^{-1} \mathbf{1})^{-1} = (n - \tau \sum_{i,j} w_{i,j})^{-1}$$

if y s are independent $\tau = 0$

When $\tau > 0$

Measures information loss in estimation of mean as a result of spatial dependence.

Dependency in spatial data

$$\text{Var}(\tilde{\mu}) = \sigma^2 (\mathbf{1}^T \mathbf{V}^{-1} \mathbf{1})^{-1}$$

So dependency effects confidence interval and significance tests

For the regression model with correlated errors

$$\beta_v = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{y})$$

$$\text{Var}(\beta_v) = \sigma^2 (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}$$

\mathbf{V} adjusts the information content in the sample.

It down weights the influence of highly correlated observations

Spatial heterogeneity: regional subdivisions and parameter variation

Data for m areal unit is available – issue is to consider different aggregation of these and estimate parameters

Selection of aggregate regions is problematic - in aggregations of regions the segmentation is not obvious

Criteria for aggregation

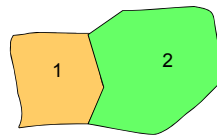
- 1) as few segments as possible,
- 2) intra -segment homogeneity
- 3) contiguous units aggregated (compactness)

Can model spatial variation in the parameters β by allowing them to vary with space

For the constant or intercept coefficient:

$$y_i = \alpha + \beta x_i + \gamma D_i + e_i$$

D_i is a dummy variable equal to 1 if site i is in area 1 and zero otherwise



For area 1 constant coefficient is $(\alpha + \gamma)$

For area 2 constant coefficient is (α)

Spatial heterogeneity: regional subdivisions and parameter variation

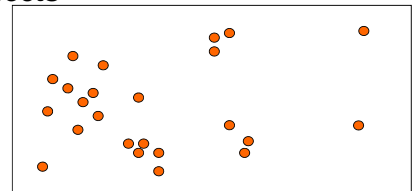
Some segmentations may not work

Large differences in numbers of observations within regions

Too few observations within regions or fewer observations than parameters to estimate

Spatial distribution of data points and boundary effects

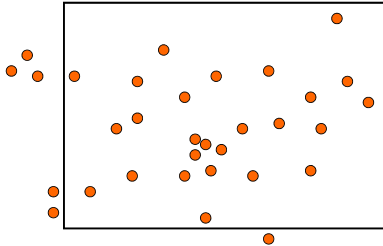
Impacts of the distribution of data points



Points with high leverage are isolated points - problems occur when these points have error in site location, or in observed response

Methods to reduce weight of clusters increases leverage of isolated points

Boundary problems in model fitting



When boundaries are arbitrary, observations close to boundary can lose neighbors

If spatial models are fit using adjacent observations, boundary neighborhoods will be incomplete

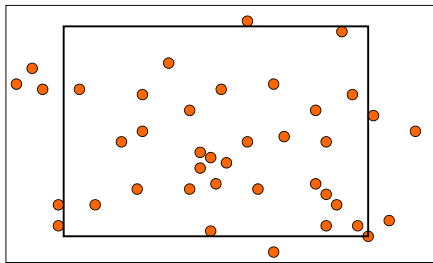
Boundary problems in model fitting

Different assumptions about unobserved boundary values lead to different model fits

Extra-regional factors can be more pronounced on boundary than interior observations

Given a model fit to the region, boundary observations may have larger variances

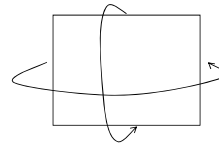
Boundary problems in model fitting



Can collect data beyond bounds of study area and fit a larger model

Boundary problems in model fitting

Model specification and estimation should examine sensitivity of results to different boundary assumptions and error assumptions



Model evaluation should distinguish between boundary and interior sites in examination of diagnostics

Isolated points next to the boundary - can have large leverage and large variance - fitting the model may need to address both effects

Assessing model fit

Part of assessing a spatial model fit is to evaluate residuals for spatial structure

Spatial structure in the residuals can indicate failure of the model to account for important elements

Need to use analysis of residual structure to help define a better model

Distributions

Non normal empirical distributions are common

Multivariate normal distributions for dependent normal data can account for correlation in variance covariance matrix

Less easy to specify correlations for other distributions (Poisson, binomial, negative binomial)

Extreme data values

Extreme data values arise in geographic problems because some areas are just different.

In fitting models these extreme values can leverage the fit such that it is misleading for the bulk of the data.

Diagnostics available for regression models with independent errors – by deletion of each observation in turn

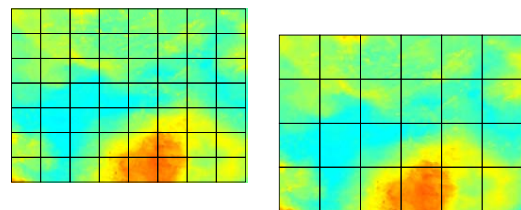
Modifications are required in case of correlated errors

Model sensitivity to areal system

Observations on areal units create distinctive problems

Consider a regular grid applied to a continuous spatial variable

Results of statistical analysis are conditional on grid size, orientation, and origin



Model sensitivity to areal system

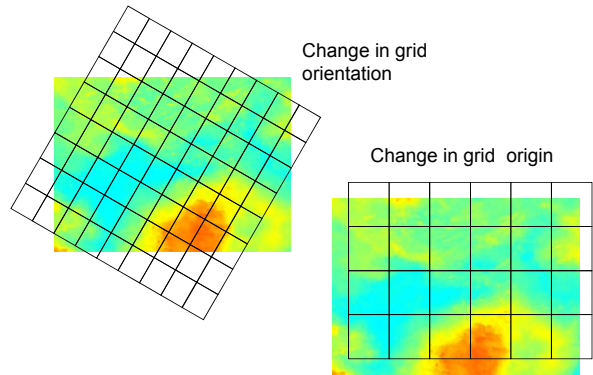
Properties of a surface smaller than the grid cell will not be detected.

Processes larger than the study area will not be detectable

Spatial surfaces have no natural (commonly accepted) origin or natural grid

Different from time series since an origin is easier to define and regular grids(units) are well established.

Model sensitivity to areal system



Severity of effect

Effects will depend on nature and scale of variation and correlation structure

Where correlation is strong in all directions and over distance up to the size of grid or grid size is small impacts not likely to be great

Irregular areal units

Primary units may be grouped arbitrarily - should try to retain intra-area homogeneity

Aggregate units should not differ greatly in size

For statistical reasons

Because processes may be operating at different scales

Model sensitivity to areal system

Results of statistical analysis will be effected by scale of study area and arrangement and scale of partitions

n areal units (observations) has different connotations

n different partitions may form many different ways and n may increase or decrease.

Changing the number of units may change the observations slightly or drastically

Modifiable areal unit problem (MAUP)

zone scale and pattern effect statistical results

makes regional comparative research difficult since zones may not be comparable

Casts doubt on results of analysis of aggregated data.

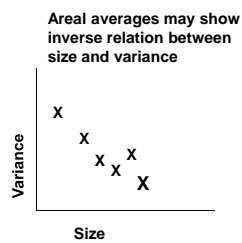
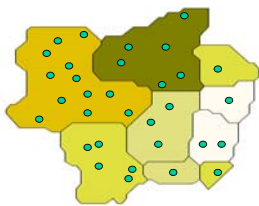
Approach

create zoning system that reflects study purpose

test several different plausible zonations for given scales

Size variance relationships in homogeneous aggregates

Issues when areal partitions contain different numbers of primary units or have very different sizes



Should investigate by plotting data values against area size

Size-variance relationships in homogeneous aggregates

Can control by weighting for large error variance in regression model

More reliable estimates receive higher weights

A Statistical Framework For Spatial Data Analysis

Exploratory data analysis - phase in which patterns and structures are uncovered and hypothesis proposed - detection of behaviors, characteristics which may cause problems for standard assumptions

Confirmatory data analysis – test hypothesis, generate confidence intervals, sensitivity analysis of model fit

Data Adaptive Modeling

