

SPATIAL STATISTICAL DATA ANALYSIS FOR GIS USERS

Konstantin Krivoruchko

CONTENTS

PREFACE

PART 1: INTRODUCTION TO STATISTICAL DATA ANALYSIS

CHAPTER 1: STATISTICAL APPROACH TO GIS DATA ANALYSIS

Spatial analysis and spatial data analysis

Types of spatial data and related statistical models

Spatial data dependency

Distance between geographic objects

An example of statistical smoothing of regional data

Assignment:

Investigate how a semivariogram can capture spatial dependence

Further reading

CHAPTER 2: EXAMPLES OF THE IMPORTANCE OF ESTIMATING DATA AND MODEL UNCERTAINTY

The difference between averaging dependent and independent data

Prediction error maps and maps that show the probability that a specified threshold is exceeded

Statistical comparison of mortality rates

The uncertainty of prediction errors

Quantile estimation and minimization of a loss function

Hypothesis testing and modeling

Assignments:

1) DEM averaging exercise

2) Create a kriging prediction error map that depends on the data values

Further reading

CHAPTER 3: UNCERTAINTY AND ERROR IN GIS DATA

Errors in GIS data

Systematic and random errors

Data variation at different scales

Using a semivariogram to detect data uncertainty

Locational uncertainty

Local data integration

Rounding-off errors

Censored and truncated data

Digital elevation model uncertainty

Case study: Error propagation in radiocesium food contamination

Estimating the internal dose in people from the measured food contamination

Estimating uncertainty in expressions with imprecise terms. Error in estimating the internal dose

Assignments:

1) Investigate the influence of the locational uncertainty on the semivariogram model

2) Recalculate the uncertainty in estimating the internal dose

Further reading

CHAPTER 4: THE IMPORTANCE OF THE DISTRIBUTION ASSUMPTION

Gaussian processes

Lognormal processes

Bernoulli, binomial, and Poisson processes

Modeling data distribution as a mixture of Gaussian distributions

Use of gamma distribution for modeling positive continuous data

Modeling proportions using beta distribution

Negative binomial distribution

Modeling data with extra zeros

Nonparametric modeling

Confidence intervals and Chebyshev's inequality

Spatial Poisson process

Assignments:

1) Simulate and plot normal, lognormal, gamma, binomial, Poisson, and negative binomial distributions

2) Detect multimodal data distributions

3) Find a golf course with the best air quality using kriging and the Chebyshev's inequality

CHAPTER 5: METHODS FOR SENSITIVITY AND UNCERTAINTY ANALYSIS

Example of sensitivity analysis in GIS

Monte Carlo simulation

Bayesian belief network

Fuzzy set theory

Fuzzy logic

Raster maps comparison

Assignments:

- 1) Use the Bayesian belief network for relating the risk of asthma to environmental factor*
- 2) Classify environmental variables using fuzzy logic*
- 3) Using fuzzy inference, find areas that most likely contain large number of people with a high irradiation dose*

Further reading

CHAPTER 6: TYPES OF SPATIAL DATA, STATISTICAL MODELS, AND MODEL DIAGNOSTICS

Three types of spatial data

Goals of spatial data modeling

Goals of spatial data exploration

Examples of applications with input data of different types: radioecology, fishery, agriculture, wine grapes quality model, wine price formation, forestry, criminology

Random variables and random fields

Stationarity and isotropy

Model diagnostic

Methods for indicator (yes/no) prediction

Methods for continuous prediction: cross-validation and validation

Summary of spatial modeling

Assignments:

- 1) Choose the appropriate model for analyzing the quantity of good-sized shoots produced by the vine*
- 2) Investigate possible models for the variation of malaria prevalence*
- 3) Calculate and display indices for indicator prediction using ozone concentration measured in June 1999 in California*
- 4) Investigate the variability of yields of individual trees*

Further reading

CHAPTER 7: SPATIAL INTERPOLATION USING DETERMINISTIC MODELS

Spatial interpolation goals

Predictions are always inaccurate

Deterministic and statistical models

Inverse distance weighted interpolation

Radial basis functions: RBFs and kriging

Global and local polynomial interpolation

Local polynomial interpolation and kriging

Interpolation using a non-Euclidean distance metric

Nontransparent barriers determined by polylines

Semitransparent barriers based on cost surface

Assignments:

1) Compare the performance of deterministic interpolation models

2) Find the best deterministic model for interpolation of cesium-137 soil contamination

Further reading

PART 2: PRINCIPLES OF MODELING SPATIAL DATA

CHAPTER 8: PRINCIPLES OF MODELING GEOSTATISTICAL DATA: BASIC MODELS AND TOOLS

Optimal prediction

Geostatistical model

Geostatistical Analyst's kriging models

Semivariogram and covariance

What functions can be used as semivariogram and covariance models?

Convolution

Semivariogram and covariance models

Models with true ranges

Powered exponential family or stable models

K-Bessel or Mattern class of covariance and semivariogram models

Models allowing negative correlations

J-Bessel semivariogram models

Rational quadratic model

Nested models

Indicator semivariogram models

Semivariogram and covariance model fitting

Trend and anisotropy

Kriging neighborhood

Continuous kriging predictions and prediction standard error

Data transformations

Data declustering

Assignments:

1) Simulate surfaces using various semivariogram models

2) Find the best semivariogram models for simulated data

3) Investigate the Geostatistical Analyst's prediction smoothing option

4) Try a general transformation of nonstationary data

Further reading

CHAPTER 9: PRINCIPLES OF MODELING GEOSTATISTICAL DATA: KRIGING MODELS AND THEIR ASSUMPTIONS

Choosing between simple and ordinary kriging

Kriging output maps

Multivariate geostatistics

Indicator kriging and indicator cokriging

Disjunctive kriging

Checking for bivariate normality

Moving window kriging

Kriging assumptions and model selection

A moving-window kriging model

Kriging with varying model parameters: sensitivity analysis and Bayesian predictions

Copula-based geostatistical models

Assignments:

1) Reproduce prediction maps shown in the "Accurate Temperature Maps Creation for Predicting Road Conditions" demo

2) Investigate the performance of simple, ordinary, and universal kriging models by comparing their predictions with known values

3) Find the optimal number of neighbors for prediction using simple and ordinary kriging models by comparing their root-mean-squared prediction errors

4) Participate in the Spatial Interpolation Comparison 97 exercise

5) Predict the tilt thickness in the lake

6) Develop a geostatistical model for interpolation of the lake Kozjak depth data

Further reading

CHAPTER 10: OPTIMAL NETWORK DESIGN AND PRINCIPLES OF GEOSTATISTICAL SIMULATION

Spatial sampling and optimal network design

Monitoring design in the precomputer and early computer era

Ideas on a network design formulated after 1963

Sequential versus simultaneous network design

Geostatistical simulation

Unconditional simulation and conditioning by kriging

Sequential Gaussian simulations

Simulating from kernel convolutions

Simulated annealing

Applications of unconditional simulations

Applications of conditional simulations

Assignments:

- 1) Find optimal places for the addition of new stations to monitor air quality in California*
- 2) Simulate a set of candidate sampling locations from inhomogeneous Poisson process*
- 3) Reduce the number of monitoring stations in the network using validation diagnostics*
- 4) Discuss two simulation algorithms proposed by GIS users that are based on estimated local mean and standard error*
- 5) Conditional simulation with Geostatistical Analyst 9.3*

Further reading

CHAPTER 11: PRINCIPLES OF MODELING REGIONAL DATA

Geostatistics and regional data analysis

The question of applying geostatistics to regional data

Binomial and Poisson kriging

Distance between polygonal features

Regional data modeling objectives

Spatial smoothing

Cluster detection methods

Spatial regression modeling

Simultaneous autoregressive model

Markov random field and conditional autoregressive model

Assignments:

- 1) Investigate the new proposal of mapping risk of disease*
- 2) Smooth the data for the tapeworm infection in red foxes*

3) Spatial clusters detection using R package DCluster

Further reading

CHAPTER 12: SPATIAL REGRESSION MODELS: CONCEPTS AND COMPARISON

Geographically weighted regression

Linear mixed model

Generalized linear model and generalized linear mixed models

Semiparametric regression

Hierarchical spatial modeling

Hierarchical models versus binomial and Poisson kriging

Multilevel and random coefficient spatial models

Geographically weighted regression versus random coefficients models

Spatial factor analysis

Copula-based spatial regression

Regional data aggregation and disaggregation

Spatial regression models diagnostics and selection

Assignments:

1) Investigate the effect of sun exposure on lip cancer deaths

2) Practice with ArcGIS 9.3 geographically weighted regression geoprocessing tool

Further reading

CHAPTER 13: PRINCIPLES OF MODELING DISCRETE POINTS

Examples of point patterns

Spatial point processes: complete spatial randomness and Poisson processes; spatial clusters; inhibition processes: Cox processes

Point pattern analysis and geostatistics

Ripley's K function

Cross K function

Pair correlation functions

Test of association between two types of point events

Model fitting

Inhomogeneous K functions

K functions on a network

Cluster analysis

Marked point patterns

Hierarchical modeling of spatial point data

Residual analysis for spatial point processes

Local indicators of spatial association

Assignments:

- 1) *Modeling the distribution of sea anemones locations*
- 2) *Modeling the distribution of early medieval grave sites*
- 3) *Getting to know Gibbs processes*

Further reading

PART 3: STATISTICAL SOFTWARE USAGE

CHAPTER 14: GEOSTATISTICS FOR EXPLORATORY SPATIAL DATA ANALYSIS

Data visualization

Exploration of ozone data clustering, dependence, distribution, variability, stationarity, and finding possible data outliers

Analysis of spatially correlated heavy metal deposition in Austrian moss

Zoning the territory of Belarus contaminated by radionuclides

Averaging air quality data in time and space

Analysis of the nonstationary data from a farm field in Illinois

Spatial distribution of thyroid cancer in children in post-Chernobyl Belarus

Assignments:

- 1) *Exploration of the arsenic groundwater contamination in Bangladesh in 1998*
- 2) *Average the particulate matter data collected in the United States in June 2002 in time and space*
- 3) *Explore the annual precipitation distribution in South Africa*

Further reading

CHAPTER 15: USING COMMERCIAL STATISTICAL SOFTWARE FOR SPATIAL DATA ANALYSIS

Programming with SAS

Traditional (nonspatial) linear regression

Linear regression with spatially correlated errors (kriging with external trend)

Using MATLAB and libraries developed by MATLAB users

Moran's I scatter plot

Simultaneous autoregressive model

Using S-PLUS spatial statistics module S+SpatialStats

Creating spatial neighbors

Moran's I

Conditional autoregressive model

Assignments:

- 1) *Repeat the Bangladesh case study using another subset of the data*
- 2) *Use MATLAB for non-Gaussian disjunctive kriging*

3) Use CAR model from S+SpatialStats module for analysis of infant mortality data collected in North Carolina from 1995–1999

Further reading

CHAPTER 16: USING FREEWARE R STATISTICAL PACKAGES FOR SPATIAL DATA ANALYSIS

Analysis of the distribution of air quality monitoring stations in California using the R splancs package

Epidemiological data analysis using the R environment and spdep package

Analysis of the relationships between two types of crime events using the splancs package

Cluster analysis using the mclust package

Assignments:

1) Simulate spatial processes with the spatstat package

2) Repeat the analysis of infant mortality using data collected in North Carolina from 1995–1999

3) Repeat the analysis of the relationships between robbery and auto theft crime events using the splancs package with 1998 Redlands data

4) Estimate the density and clustering of gray whales near the coastline of Flores Island 5) Test for the spatial effects around putative source of health risk

Further reading

APPENDIX 1: USING ARCGIS GEOSTATISTICAL ANALYST 9.2

Exploratory spatial data analysis

Displaying the semivariogram surface on the map

Statistical predictions

Predictions using ordinary kriging

Replicated data prediction using lognormal ordinary kriging

Continuous predictions

A close look at predictions with replicated data

Quantile map creation using simple kriging

Multivariate predictions

Probability map creation using cokriging of indicators

Probability maps creation using disjunctive and ordinary kriging

Moving window kriging

Example of geoprocessing: finding places for new monitoring stations

Deterministic models

Validation diagnostics

About ArcGIS Geostatistical Analyst 9.3

Assignments:

- 1) Repeat the analysis shown in this appendix*
- 2) Use the Geostatistical Analyst models and tools to analyze heavy metals measurements collected in Austria in 1995*
- 3) Find the 20 best places for collecting new values of arsenic to improve predictions of this heavy metal distribution over Austrian territory*
- 4) Practice with the Gaussian Geostatistical Simulation geoprocessing tool*

Further reading

APPENDIX 2: USING R AS A COMPANION TO ARCGIS

Downloading

The first R session

Reading and displaying the data

Scatterplots

The linear models

Fitting a linear model in R

Regression diagnostics

Beyond linear models

Running R scripts from ArcGIS

Assignments:

- 1) Repeat linear regression analysis of data on infant mortality data and house prices*
- 2) Verify the assumptions of the linear regression model*

Further reading

APPENDIX 3: INTRODUCTION TO BAYESIAN MODELING USING WINBUGS

On the reasons for using Bayesian modeling

Bayesian regression analysis of housing data

Multilevel Bayesian modeling

Bayesian conditional geostatistical simulations

Regional Bayesian analysis of thyroid cancer in children in Belarus

Assignments:

- 1) Repeat the case studies presented in this appendix*
- 2) Interpolate the precipitation data using Gaussian and Bayesian kriging*
- 3) Perform the Bayesian analysis of weeds data*
- 4) Verify the classification of the happiest countries in Europe*
- 5) Bayesian random coefficient modeling of the crime data*
- 6) Bayesian spatial factor analysis of the crime data*

Further reading

APPENDIX 4: INTRODUCTION TO SPATIAL REGRESSION MODELING USING SAS

Logistic regression

Logistic regression with spatially correlated errors

Poisson regression with spatially correlated errors

Binomial regression with spatially correlated errors

Semivariogram modeling using Geostatistical Analyst and the procedures mixed and nlin

Assignments:

1) Repeat analysis of the pine beetle and thyroid cancer data

2) Reconstruct the semivariogram model parameters

3) Compare two semivariogram models fitting

Further reading

AFTERWORD

GLOSSARY

BIBLIOGRAPHY