Geospatial statistics and spatial data interpolation methods

ERASMUS Intensive Program
GIS’EM 2013 at Eberswalde

Prof. Dr. Jan-Peter Mund
University for Sustainable Development, Eberswalde
Spatial data interpolation
Outline - Exploring Continuous Data

- Spatial Interpolation
- Sampling
- Triangulation
- Inverse Distance Weighting (IDW)
- Trend
- Spline
- Kriging
Desired Learning Objectives

- Define spatial interpolation
- Describe the inputs needed for spatial interpolation
- List global and local interpolation methods;
  - Describe the differences between the 2 methods
- Describe the difference between an exact and an inexact interpolation method
- Define kriging
  - How does it differ from other interpolation methods?
- Define sampling methodology
  - Describe applications of various sampling methodologies
Essential skill analysing environmental data

- Environmental data
  - often collected as discrete observations at points or along transects
  - example: soil cores, soil moisture, vegetation transects, meteorological station data, etc.

- Need to convert discrete data into continuous surface for use in GIS modelling
  - interpolation
Spatial Interpolation

- The process of using points with known values to estimate values at other (often unknown) points

- Typically applied to a raster
  - We estimate values for all cells

- Therefore, spatial interpolation is a means of creating surface data from sample points!
How is it used...
Which One’s the DEM?

(a) Discrete Elevation Samples

(b) Implicit (Linear) Continuous Surface
Introduction

- **Definition:**
  
  “Spatial interpolation is the procedure of estimating the values of properties at unsampled sites within an area covered by existing observations.”
  
  (Waters, 1989)

- **Complex problem**
  
  - wide range of applications
  - important in addressing problem of data availability
  - quick fix for partial data coverage
  - interpolation of point data to surface/polygon data
  - role of filling in the gaps between observations
What is spatial analysis?

- Methods for working with spatial data
  - to detect patterns, anomalies
  - to find answers to questions
  - to test or confirm theories
    - deductive reasoning
  - to generate new theories and generalizations
    - inductive reasoning
- "a set of methods whose results change when the locations of the objects being analyzed change"
What is interpolation?

- Process of creating a surface based on values at isolated sample points.
- Sample points are locations where we collect data on some phenomenon and record the spatial coordinates.
- We use mathematical estimation to “guess at” what the values are “in between” those points.
- We can create either a raster or vector interpolated surface.
- Interpolation is used because field data are expensive to collect, and can’t be collected everywhere.
What isn’t interpolation?

- Interpolation only works where values are spatially dependent, or spatially autocorrelated, that is, where nearby location tend to have similar Z values.

- Examples of spatially autocorrelated features: elevation, property value, crime levels, precipitation

- Non-autocorrelated examples: number of TV sets per city block; cheeseburgers consumed per household.

- Where values across a landscape are geographically independent, interpolation does not work because value of \((x,y)\) cannot be used to predict value of \((x+1, y+1)\).
Where interpolation does NOT work

- Cannot use interpolation where values are not spatially autocorrelated

- Say looking at household income—in an income-segregated city, you could take a small sample of households for income and probably interpolate

- However, in a highly income-integrated city, where a given block has rich and poor, this would not work
Spatial data sets can be ...

- **Dense**
  - Many data points per hectare
  - e.g., grain yield data sets often consist of 300 to 600 data points per hectare

- **Sparse**
  - Fewer data points per hectare
  - e.g., typical grid soil sampling results in an average of 0.4 data point per hectare
Yield data are dense …

- One sec. readings at 1-100 data point every meter
  - 600-3000 data points per hectare with a 6-row combine header
Soil sample data are sparse

- **Typical 1 hectare sampling grid**
  - Only 6-12 point per hectare
Organic matter surface map

- Interpolated from values of 1 ha soil sample data
Reality check

- Soil surface color from reclassified aerial IR
- Soil surface map interpolated from 1 ha samples
Data sampling

- Method of sampling is critical for subsequent interpolation...

- Regular
- Random
- Transect
- Stratified random
- Cluster
- Contour
Sampling

- Systematic sampling pattern
  - Easy
  - Samples spaced uniformly at fixed X, Y intervals
  - Parallel lines

- Advantages
  - Easy to understand

- Disadvantages
  - All receive same attention
  - Difficult to stay on lines
  - May be biases
Sampling

Random Sampling

- Select point based on random number process
- Plot on map
- Visit sample

Advantages

- Less biased (unlikely to match pattern in landscape)

Disadvantages

- Does nothing to distribute samples in areas of high
- Difficult to explain, location of points may be a problem
Sampling

- **Cluster Sampling**
  - Cluster centers are established *(random or systematic)*
  - Samples arranged around each center
  - Plot on map
  - Visit sample
    - *(e.g. Forest Services, Forest Inventory Analysis (FIA)*
    - *Clusters located at random then systematic pattern of samples at that location)*

- **Advantages**
  - Reduced travel time
Sampling

- **Adaptive sampling**
  - Higher density sampling where the feature of interest is more variable.
  - Requires some method of estimating feature variation
  - Often repeat visits (e.g. two stage sampling)

- **Advantages**
  - Often efficient as large homogeneous areas have few samples reserving more for areas with higher spatial variation.

- **Disadvantages**
  - If no method of identifying where features are most variable then several you need to make several sampling visits.
Types of Spatial Interpolation

- Spatial interpolation can be **global** or **local**
- Spatial interpolation can be **exact** or **inexact**
- Spatial interpolation can be **deterministic** or **stochastic**
Interpolation Techniques

- Thiessen Polygons
- TINS
- Inverse Distance Weighting (IDW)
- Trend
- Spline
- Kriging
Simple Triangulation

Tree height = \tan (30^\circ) \times 50 \text{ m} = 28.86 \text{ m}

Angle (degrees) = 30^\circ
Angle (percent) = \left(\frac{28.86 \text{ m}}{50 \text{ m}}\right) = 57.7\%

tan (30^\circ) 100 = 57.7, providing a quick conversion from degrees to percent slope

A simple example of the conversion process from degrees to percent slope.
Visibility process

- Viewpoints (observers) are often point locations
- Operator can set limits (distance, angle, height)
- Line of sight from a viewing site to the surrounding landscape
Wind farm – photomontage

before

wind farm CAD frame model

after
Spatial data interpolation – Principles

Concepts and principles
Spatial data interpolation – Principles

? = unknown elevation
★ = known elevations

Distance

200'
★
100'
50'
350'
70'
300'

250'
★
150'
Interpolation for Visibility Analysis

- What’s the profile through the “cell”? 

![Diagram showing interpolation for visibility analysis with points A, B, C, and D, and grid coordinates 50, 54, 68, and 72.](image)
Surfaces from points

Points

Surface
Control Points

- Control Points (CPs) are points with known values
  - CPs provide the data necessary for the development of an interpolator for spatial interpolation
- The *number and distribution* of control points can greatly influence the *accuracy* of spatial interpolation.
Exact vs Inexact Interpolation

Exact interpolation goes thru known values!

(a)

(b)
DEM production with a long historical tradition

Reverse side of tetradrachm coin representing the pattern of major massifs of the region of Ephesus (issued between 336-334 BC to pay the Persian army)

The same region as depicted by the DEM/SRTM in 2004
TIN construction

Plan view

Isometric view

Interpolated value x

value a

value b

value c

a

b

c
DEM and TINs

TIN based on same sample points

DEM with sample points
Delaunay Triangulation

- Two approaches
  - Define triangles by finding three points that define a circle that doesn’t include any other points
  - Tessellate by assigning all locations to the nearest vertex
    - Boundaries form a set of polygons called Thiessen or Voronoi polygons
Raster DEM

- Tessellation of regular triangles, squares and hexagons
- Simple derivates, filtering, ...

On the left is a point dataset of known values. On the right is a raster interpolated from these points. Unknown values are predicted with a mathematical formula that uses the values of nearby known points.
Deterministic vs. Stochastic Interpolation

- Deterministic: Does *not* provide an estimate of errors associated with predicted values (no statistics!)

- Stochastic: Provides and assessment of prediction errors with estimated variances (statistics as a "measure of goodness" of the interpolation)
Classifying Spatial Interpolation Methods

<table>
<thead>
<tr>
<th>Global</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic</td>
<td>Deterministic</td>
</tr>
<tr>
<td>Trend surface (inexact)*</td>
<td>Thiessen (exact)</td>
</tr>
<tr>
<td>Regression (inexact)</td>
<td>Density estimation (inexact)</td>
</tr>
<tr>
<td></td>
<td>Inverse distance weighted (exact)</td>
</tr>
<tr>
<td></td>
<td>Splines (exact)</td>
</tr>
<tr>
<td>Stochastic</td>
<td>Stochastic</td>
</tr>
<tr>
<td>Regression (inexact)</td>
<td>Kriging (exact)</td>
</tr>
</tbody>
</table>

*Given some required assumptions, trend surface analysis can be treated as a special case of regression analysis and thus a stochastic method (Griffith and Amrhein 1991).
Does my data follow a normal distribution?

- What should I look for?
  - Bell-shaped
  - No outliers
  - Mean ≈ Median
  - Skewness ≈ 0
  - Kurtosis ≈ 3
Does my data follow a normal distribution?
What is autocorrelation?

Tobler’s first law of geography:
"Everything is related to everything else, but near things are more related than distant things."

![Graphs showing autocorrelation](image)
Spatial Interpolation in GIS

- Estimation of z-value for any point location within a specific spatial area

Assume:

- The data is continuous
- The data are spatially dependent; it can be estimated based on surrounding locations
Spatial Interpolation

- **Continuous Data**
  - **Whole Area Methods** – interpolation based on all points in a study area
    - Trend Surface Analysis
    - Fourier Series
  - **Local Interpolators** – interpolation may be applied to only a portion of the data
    - Moving Average
    - Splines
    - Kriging
Inverse Distance Weighted

- Each input point has local influence that diminishes with distance
- estimates are averages of values at n known points within window

\[ z(x) = \frac{\sum_i w_i z_i}{\sum_i w_i} \]

- where \( w \) is some function of distance

\[ w_i = \frac{1}{d_i^2} \]
Inverse Distance Weighting (IDW)

\[ Z_j = \frac{\sum_{i} \frac{Z_i}{d_{ij}^n}}{\sum_{i} \frac{1}{d_{ij}^n}} \]

- \( Z_i \) is value of known point
- \( D_{ij} \) is distance to known point
- \( Z_j \) is the unknown point
- \( n \) is a user selected exponent (often 1, 2 or 3)

Any number of points may be used up to all points in the sample; typically 3 or more.
Inverse Distance Weighting

In Geostatistical Wizard, IDW Interpolation: Step 1 of 2 - Set Parameters:
- Symbol Size: 40%
- Method: Neighborhood
- Neighbors to Include: 15
- Include at Least: 10
- Shape Type: Square
- Shape Angle: 0°
- Major Semiaxis: 4324
- Minor Semiaxis: 4324
- Anisotropy Factor: 1

In Geostatistical Wizard, IDW Interpolation: Step 2 of 2 - Cross Validation:
- Chart:
  - Regression function: 0.446 x + 1.750
  - Prediction Errors:
    - Mean: 0.04331
    - Root-Mean-Square: 2.639
  - Samples: 852 of 852

Prof. Dr. rer. nat. Jan-Peter Mund
GIS and NRM
HNE Eberswalde (FH) 06/03/2013
Inverse Distance Weighting
A potentially undesirable characteristic of IDW interpolation

- This set of six data points clearly suggests a hill profile.
- But in areas where there is little or no data the interpolator will move towards the overall mean.
- Blue line shows the profile interpolated by IDW
Whole Area Interpolators

- Trend Surface
  - When an order higher than 1 is used, the interpolator may generate a Grid whose minimum and maximum might exceed the minimum and maximum of the input points.
  - As the order of the polynomial is increased, the surface being fitted becomes progressively more complex.
- Main Use
  - Not an interpolator within a region, but a way of removing broad features of the data prior to using some other local interpolator.
Fitting a single polynomial trend surface

- interpolated point
- data point
Example trend surfaces

Source surface with sample points

Linear
Goodness of fit (R2) = 45.42 %

Quadratic
Goodness of fit (R2) = 82.11 %

Cubic
Goodness of fit (R2) = 92.72 %
Global Polynomial Interpolation

- Global polynomial interpolation technique fits a plane through the measured data points. A plane is typically a polynomial.
Global Polynomial Interpolation

![Global Polynomial Interpolation Figure]

- Layers: 
  - GPItestPredictions
  - Global Polynomial Interpolation
    - Prediction Map
      - Observed training
      - Observed_test
      - county_Union1

![ArcMap Interface]

- File Edit View Insert Selection Tools Window Help
- Editor
- Task: Create New Feature
- Target:
- 1,364,352

---

GIS and NRM

HNE Eberswalde (FH) 06/03/2013
Local polynomial Interpolation

- While Global Polynomial interpolation fits a polynomial to the entire surface, Local Polynomial interpolation fits many polynomials, each within specified overlapping neighborhoods.
Spatial moving average (SMA)

\[
\frac{3 + 7 + 5}{3} = 5
\]

\[
\frac{3 + 7 + 5}{3} = 5
\]
Example SMA (circular filter)

Source surface with sample points

11x11 circular filter SMA with sample points
21x21 circular filter SMA
41x41 circular filter SMA
Evaluation Criteria

- Several statistical indicators (Root Mean Square Error (RMSE), Mean Error (ME), Mean Absolute Error (MAE) and Mean Square Error (MSE)) are computed on observed and predicted radon concentrations.

- Confidence limits on the statistics for
  - Normalized Mean Square Error (NMSE),
  - Fractional Bias (FB), and
  - Coefficient of Correlation (r)

- are calculated using Bootstrap application to identify the most suitable interpolation technique.
Spline Method

Another option for interpolation method:

- This fits a curve through the sample data assign values to other locations based on their location on the curve.
- Thin plate splines create a surface that passes through sample points with the least possible change in slope at all points, that is with a minimum curvature surface.
- SPLINE has two types: regularized and tension.
- Tension results in a rougher surface that more closely adheres to abrupt changes in sample points.
- Regularized results in a smoother surface that smoothes out abruptly changing values somewhat.
Splines

- Name derived from the drafting tool, a flexible ruler, that helps create smooth curves through several points
- Spline functions (also called splines) are used to interpolate along a smooth curve. (similar to the flexible ruler).
- Fits a minimum-curvature surface through the input points. Conceptually, it is like bending a sheet of rubber to pass through the points, while minimizing the total curvature of the surface
- Force a smooth line to pass through a desired set of points
- Constructed from a set of joined polynomial functions
Kriging

- Similar to Inverse Distance Weighting (IDW)
- Kriging uses the minimum variance method to calculate the weights rather than applying an arbitrary or less precise weighting scheme
- A set of sample points are used to estimate the shape of the Variogram model is made
- A line is fit through the set of semi-variance points
- The Variogram model is then used to interpolate the entire surface
Kriging assumptions

- Some spatial surfaces cannot be modeled using deterministic methods that use smooth mathematical functions.
- Specifically if data are sparse, for example ground-water modeling, gravity data, soil mapping, water toxicity, air pollution, bathymetric data etc.
- Kriging is a **stochastic** interpolation method in contrast with **deterministic** methods (TIN, Inverse distance, trend estimation)
- It attempts to statistically obtain the **optimal prediction** i.e. to provide the Best Linear Unbiased Estimation (BLUE), specifically when **data are sparse**
Kriging

- Lag distance
- Where:
  - Zi is a variable at a sample point
  - hi is the distance between sample points
- Every possible set of pairs Zi, Zj defines a distance hij, and is different by the amount
  - Zi – Zj.
- The distance hij is known as the lag distance between point i and j. Also there is a subset of points in a sample set that are a given lag distance apart

Figure 9-12: Lag distances, used in calculating semi-variances for kriging.
Kriging

- Semi-variance

\[ \gamma(h) = \frac{1}{2n} \sum (Z_i - Z_j)^2 \]

- Where \( Z_i \) is the measured variable at one point
- \( Z_j \) is another at \( h \) distance away
- \( n \) is the number of pairs that are approximately \( h \) distance apart

- Semi-variance may be calculated for any \( h \)
- (When nearby points are similar (\( Z_i - Z_j \)) is small so the semi-variance is small. High spatial autocorrelation means points near each other have similar \( Z \) values)
Ordinary Kriging

There are three primary parameters that describe the autocorrelation of radon concentrations. These are range, nugget and sill.

- The range is where the best-fit line starts to level off, (46.55). Within the range, all data are correlated.

- The maximum semivariogram value is sill parameter (0.2869)

- Nugget is data variation due to measurement errors (0.20487).
Ordinary Kriging
Ordinary Kriging
Analysis of Variogram
How Many Neighbors?

Geostatistical Wizard: Step 3 of 4 - Searching Neighborhood

Dataset Selection: Dataset 1

Symbol Size: 3

Method: Neighborhood

Neighbors to Include: 40

Include at Least: 2

Shape Type:

Angle: 0.0

Major Semiaxis: 168410

Minor Semiaxis: 168410

Anisotropy Factor: 1

Test Location:

X: -2048573.4

Y: 212235.76

Neighbors: 26

Prediction: 0.10641

Preview type: Neighbors

< Back  Next >  Finish  Cancel
Example

- Here are some sample elevation points from which surfaces were derived using the three methods.
Example: IDW

- Done with $P = 2$. Notice how it is not as smooth as Spline. This is because of the weighting function introduced through $P$. 

![IDW Example](image_url)
Example: Spline

- Note how smooth the curves of the terrain are; this is because Spline is fitting a simply polynomial equation through the points.
Example: Kriging

- This one is kind of in between—because it fits an equation through point, but weights it based on probabilities
Inverse Distance Weighting
Kriging
### A Comparison of the Geostatistical Analyst methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Deterministic/ Stochastic</th>
<th>Output Surface Types</th>
<th>Computing Time/Modeling Time</th>
<th>Exact Interpolator</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Assumptions²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inverse Distance Weighted</td>
<td>Deterministic</td>
<td>Prediction</td>
<td>Fast/Fast</td>
<td>Yes</td>
<td>Few parameter decisions</td>
<td>No assessment of prediction errors; produces “bulls eyes” around data locations</td>
<td>None</td>
</tr>
<tr>
<td>Global polynomial</td>
<td>Deterministic</td>
<td>Prediction</td>
<td>Fast/Fast</td>
<td>No</td>
<td>Few parameter decisions</td>
<td>No assessment of prediction errors; may be too smooth; edge points have large influence</td>
<td>None</td>
</tr>
<tr>
<td>Local polynomial</td>
<td>Deterministic</td>
<td>Prediction</td>
<td>Moderately Fast/Moderate</td>
<td>No</td>
<td>More parameter decisions</td>
<td>No assessment of prediction errors; may be too automatic</td>
<td>None</td>
</tr>
<tr>
<td>Radial basis functions</td>
<td>Deterministic</td>
<td>Prediction</td>
<td>Moderately Fast/Moderate</td>
<td>Yes</td>
<td>Flexible and automatic with some parameter decisions</td>
<td>No assessment of prediction errors; may be too automatic</td>
<td>None</td>
</tr>
<tr>
<td>Kriging</td>
<td>Stochastic</td>
<td>Prediction; Standard Errors; Probability; Quantile</td>
<td>Moderately Fast/Slower</td>
<td>Yes without measurement error; No with measurement error</td>
<td>Very flexible; allows assessment of spatial autocorrelation; can obtain prediction standard errors; many parameter decisions</td>
<td>Need to make many decisions on transformations, trends, models, parameters, and neighborhoods</td>
<td>Data comes from a stationary stochastic process, and some methods require that the data comes from a normal distribution</td>
</tr>
<tr>
<td>Cokriging</td>
<td>Stochastic</td>
<td>Prediction; Standard Errors; Probability; Quantile</td>
<td>Moderate/Slowest</td>
<td>Yes without measurement error; No with measurement error</td>
<td>Very flexible; can use information in multiple datasets; allows assessment of spatial cross-correlation; many parameter decisions</td>
<td>Need to make many decisions on transformations, trends, models, parameters, and neighborhoods</td>
<td>Data comes from a stationary stochastic process, and some methods require that the data comes from a normal distribution</td>
</tr>
</tbody>
</table>

1. **Computing time** is computer-processing time to create a surface. **Modeling time** includes user-processing time to make decisions on model parameters and search neighborhoods.
2. We assume that all methods are predicting a smooth surface from noisy data.
Which Method to Use?

- Trend - rarely goes through your original points
- Spline - best for surfaces that are already smooth
  - Elevations, water table heights, etc.
- IDW - assumes variable decreases in influence w/ distance from sampled location
  - Interpolating a surface of consumer purchasing power for a retail store
- Kriging - if you already know correlated distances or directional bias in data
  - Geology, soil science
Which to Use? cont.

- Kriging - Allows user greater flexibility in defining the model to be used in the interpolation
  - Tracks changes in spatial dependence across study area (may not be linear)
  - Produces
    - a smooth, interpolated surface
    - variogram (how well pixel value fits overall model)
      - Diagnostic tool to refine model
      - Want to get variances close as possible to zero
Interpolation Software

- ArcGIS with Geostatistical Analyst
  - ArcView 3.3
- Surfer (Golden Software)
  - Surface II package (Kansas Geological Survey)
  - GEOEAS (EPA)
  - Spherekit (NCGIA, UCSB)
- Matlab
Effects of data uncertainty

Original surface

Interpolation based on 100 points

Error map

Interpolation based on 10 points

Error map

- Low
- Medium
- High
Edge effects

Original surface with sample points
Interpolated surface
Error map and extract

- Low
- Blue
- Yellow
- High
Conclusions

- Interpolation of environmental point data is important skill
- Many methods classified by
  - local/global, approximate/exact, gradual/abrupt and deterministic/stochastic
  - choice of method is crucial to success
- Error and uncertainty
  - poor input data
  - poor choice/implementaiton of interpolation method
Conclusion

- Prediction maps were created using the training data set for all five interpolation techniques and projected values were estimated for the test data set.

- Statistical parameters (error values) were evaluated and the prediction maps generated from these techniques were compared to the soil uranium concentration map.

- It was inferred that any of the four (Ordinary Kriging, IDW, RBF and Local Polynomial) interpolation techniques can be used for predicting the radon concentrations for unmeasured zip codes.

- Ordinary Kriging technique was chosen and the geometric means of radon concentrations were evaluated for unmeasured zip codes.
Literature

New in Geostatistical Analyst 10.1 : Areal Interpolation

by Eric Krause on June 8, 2012

For version 10.1, we’ve taken on a classic problem in GIS: how to reallocate data from one set of polygons to a different set of polygons. For example, demographers frequently collect data from various sources, so their data might be a mixture of census block groups, postal codes, and county boundaries. However, to perform an accurate analysis, they might need all of their data in the same administrative units.

While there are various methods for going from small polygons to large polygons (from census blocks to postal codes, for example), the benefit of areal interpolation is that it additionally provides a statistically accurate framework for going from large polygons to small polygons. By convention, the starting polygons are called the “source” polygons, and the ending polygons are called the “target” polygons. Continue reading →

Posted in Analysis & Geoprocessing | Tagged Geostatistical Analyst, interpolation, polygon | 3 Comments

New in Geostatistical Analyst 10.1 : Empirical Bayesian Kriging

by Eric Krause on June 8, 2012

Those of you familiar with kriging interpolation know that it is not always the easiest technique to implement successfully. For a long time we’ve wanted to make a geoprocessing tool that can automate kriging, but the problem has always been in the complexity of calculating good default parameters. At 10.1, through a combination of subsetting and simulations, we have a solution to the problem with a method called empirical Bayesian kriging (EBK). The method is available in the Geostatistical Wizard and as a geoprocessing tool in the Geostatistical Analyst toolbox.

EBK works by building local models on subsets of the data, which are then combined together to create the final surface. Because the interpolation model is built automatically, the method requires very few parameters. There are also some optional parameters that give you some control over how locally the models will be built and how they will be combined together.

Why should I use EBK?

- Simplicity – To get accurate results, all you need to do is specify the field you want to interpolate. Other kriging methods require you to build the model step-by-step to be confident that the results are statistically accurate.
- Automation – Because EBK is available as a geoprocessing tool, you can use it in Model Builder and in Python scripts.
- Capture small-scale effects – Using local models allows EBK to capture small-scale effects that global kriging models may miss.

This post was contributed by Eric Krause, a product engineer on the analysis and geoprocessing team.
Thank you for your kind attention