Some Geostatistical Concepts for Interoperability

Phaedon C. Kyriakidis

Department of Geography
University of California Santa Barbara

www.geog.ucsb.edu/~phaedon
Scale Issues in Interoperability

A common problem:

- too often, data come in different frames than those required for analysis
- pressing need to transform data across frames, in a rigorous way

Some (unfortunately) common solutions:

- data of disparate resolutions often treated as point measurements
- spatial correlation and data quality are inadequately accounted for
- internal consistency of conflated predictions not guaranteed
- uncertainty in conflated predictions not reported

In this talk, a unifying framework for overcoming such problems
**Support**

- **Definition:** support = domain informed by each datum or unknown value

- **Source supports:** domains over which integrated measurements are obtained

  \[ \Omega_s = \{ s_k, k = 1, \ldots, K \} \]

  \[ |s_k| = \text{measure (length, area, volume) of } k\text{-th source support} \]

- **Target supports:** domains over which predictions are sought after

  \[ \Omega_t = \{ t_p, p = 1, \ldots, P \} \]

  \[ |t_p| = \text{measure (length, area, volume) of } p\text{-th target support} \]

  Supports can be completely disjoint, or partially overlapping, but not coincident (source) with otherwise arbitrary shape, size, and orientation
Examples of Point Supports

Spot heights or transects

Regular grid

TIN

Contour lines

+ Inequality constraints
Examples of Non-Point Supports

*filled polygons = source supports*

*empty polygons = target supports*

**Note:** Point support = limiting case of non-point support
Underlying (Latent) Point Support Field

\[ \{ z(\mathbf{x}), \mathbf{x} \in D \} \]

\( \mathbf{x} = \) coordinate vector (could be defined in 4D) \( z(\mathbf{x}) = \) attribute value at \( \mathbf{x} \)

Characteristics:

- actual values are unknown, apart from some point sample data (if any)
- all viewed as realizations of a stationary random field (RF) model
- which is parameterized by a mean and covariance function
- non-stationarity, as well as auxiliary variables, can also be accounted for
Statistics of Latent Point Field

**Stationary mean:** \( E \{ Z(\mathbf{x}) \} = m_Z, \forall \mathbf{x} \)

same for all locations, typically unknown

**Stationary covariance:** \( \text{Cov} \{ Z(\mathbf{x}), Z(\mathbf{x}') \} = C_Z(\mathbf{x} - \mathbf{x}') \)

function only of distance (and possibly orientation) of vector joining \( \mathbf{x} \) and \( \mathbf{x}' \)

Shape of covariance characterizes smoothness of latent point field

![Covariance functions](image)
Source Data

\[ z(s_k) = \int_{x \in s_k} g_k(x)z(x) \, dx, \quad k = 1, \ldots, K \]

- \( z(s_k) \) = known integrated measurement of unknown point values in support
- \( g_k(x) \) = known kernel quantifying how points contribute to \( k \)-th source datum

\( g_k(x) = 1, \) if \( x \in s_k, 0 \) if not \hspace{1cm} \( g_k(x) = 1/|s_k|, \) if \( x \in s_k, 0 \) if not

- spatial statistics of source data functionally linked to those of latent point field

for extensive variables \hspace{1cm} for intensive variables

for irregular supports, even if the point field is stationary, the statistics (mean covariance) of source data are not
Statistics of Source Data

Mean of $k$-th source support: $m_Z(s_k) = E\{Z(s_k)\} = m_Z \int_{x \in S_k} g_k(x) \, dx$

Non-stationary,
unless all source supports and all source sampling functions are the same

Auto-covariance between two source supports

$$C_Z(s_k, s_k') = \text{Cov}\{Z(s_k), Z(s_k')\} = \int_{x \in S_k} g_k(x) \int_{x' \in S_k'} g_k'(x') \, C_Z(x - x') \, dx' \, dx$$
\[ z(t_p) = \int_{x \in t_p} g_p(x) z(x) \, dx, \quad p = 1, \ldots, P \]

- \( z(t_p) = \text{unknown} \) integrated value of \( \text{unknown} \) point values in support

- \( g_p(x) = \text{known} \) kernel quantifying how points contribute to \( k \)-th target value

- Statistics of target values functionally linked to those of latent point field

*for irregular supports, even if the point field is stationary, the statistics (mean covariance) of target values are not*
Statistics of Target Values

Mean of $p$-th target support:  
\[ m_Z(t_p) = E \{Z(t_p)\} = m_Z \int_{x \in t_p} g_p(x) \, dx \]

Non-stationary,  
unless all target supports and all target sampling functions are the same

Variance of a target support  
\[ C_Z(t_p, t_p) = \text{Var} \{Z(t_p)\} = \int_{x \in t_p} g_p(x) \int_{x' \in t_p} g_p(x') \, C_Z(x - x') \, dx' \, dx \]

Non-stationary

point covariance
Linking Target Values and Source Data

Cross-covariance between target and source supports:

\[ C_Z(t_p, s_k) = \text{Cov}\{Z(t_p), Z(s_k)\} = \int_{x \in t_p} g_p(x) \int_{x' \in s_k} g_k(x') C_Z(x - x') \, dx' \, dx \]

Cross-covariance between target support and source point:

\[ C_Z(t_p, x_k) = \text{Cov}\{Z(t_p), Z(x_k)\} = \int_{x \in t_p} g_p(x) C_Z(x - x_k) \, dx \]
Predicting Target Values

Assuming a known point mean \( m_z = 0 \), for simplicity

\[
\hat{z}(t_p) = w_p^T z_s = \begin{bmatrix}
w_p(s_1) & \cdots & w_p(s_K)
\end{bmatrix}
\begin{bmatrix}
z(s_1) \\
\vdots \\
z(s_K)
\end{bmatrix}
\]

Recall that some source data might be of point support

**System of Normal Equations for finding weights:**

\[
\begin{bmatrix}
C_Z(s_1,s_1) & \cdots & C_Z(s_1,s_K) \\
\vdots & \ddots & \vdots \\
C_Z(s_K,s_1) & \cdots & C_Z(s_K,s_K)
\end{bmatrix}
\begin{bmatrix}
w_p(s_1) \\
\vdots \\
w_p(s_K)
\end{bmatrix}
= \begin{bmatrix}
C_Z(t_p,s_1) \\
\vdots \\
C_Z(t_p,s_K)
\end{bmatrix}
\]

Unique solution exists, as long as point covariance model is positive definite
Uncertainty in Target Predictions

**Kriging weights:**

\[
\begin{pmatrix}
    w_p(s_1) \\
    \vdots \\
    w_p(s_K)
\end{pmatrix} =
\begin{pmatrix}
    C_Z(s_1,s_1) & \ldots & C_Z(s_1,s_K) \\
    \vdots & \ddots & \vdots \\
    C_Z(s_K,s_1) & \ldots & C_Z(s_K,s_K)
\end{pmatrix}^{-1}
\begin{pmatrix}
    C_Z(t_p,s_1) \\
    \vdots \\
    C_Z(t_p,s_K)
\end{pmatrix}
\]

Account for correlation between target and source data, after the latter have been discounted for their redundancy

**Minimum variance of prediction error**

\[
\sigma(t_p) = C_Z(t_p,t_p) - w_p \Sigma_p^{ts} = C_Z(t_p,t_p) - \begin{bmatrix}
    w_p(s_1) & \ldots & w_p(s_K)
\end{bmatrix}
\begin{bmatrix}
    C_Z(t_p,s_1) \\
    \vdots \\
    C_Z(t_p,s_K)
\end{bmatrix}
\]

Minimum among all other weighted linear combinations (for given point covariance model)
Downscaling Example

Reference point values

Pixel support data

averages of point values over pixels of size = 10x10

Point covariance #1

mass-preserving downscaled predictions

Point covariance #2

white noise covariance
No Spatial Correlation Solution

For non-overlapping source supports

White noise point covariance model:
\[
C_Z(x - x') = C_Z(0) \delta_{xx'}
\]
\[
\delta_{xx'} = 1, \text{ if } x = x', 0 \text{ if not}
\]
\[
C_Z(0) = \text{Var}\{Z(x)\}
\]

Non-zero weights only for source supports that overlap with target

\[
w_p(s_k) = \frac{C_Z(t^o_p, s_k)}{C_Z(s_k, s_k)} = \frac{\int_{x \in t_p} g_p(x) g_k(x) \, dx}{\int_{x \in s_k} g_k(x) g_k(x) \, dx}
\]

for extensive variables

\[
\left| \frac{t_p \cap s_k}{|s_k|} \right|
\]

for intensive variables

Area-weighting scheme

\[
w_p(s_k) = \left| \frac{t_p \cap s_k}{|t_p|} \right|
\]
Change of Support Recap

Framework for transforming data on a spatial frame to another:

- **general**: can handle integrated measurements over arbitrary domains
- **simple**: utilizes standard geostatistical theory with minor modifications
- **comprehensive**: can handle alternative types of point covariance models
- **consistent**: guarantees reproduction of data at larger scales (mass preserving)
- **providing uncertainty assessment**: regarding target predictions

*Point covariance model is actually a free parameter = inference problem?*

*Proposed framework just provides multiple consistent solutions to Ecological Fallacy*
Change of Support Extensions

Accounting for auxiliary variables:

- available at both source and target supports: via generalized least squares regression (**Universal Kriging**)
- available at arbitrary supports: via coKriging

Accounting for non-Gaussian data (**heteroscedastic prediction error variance**):

- Lognormal or Disjunctive Kriging

Accounting for uncertainty in source data:

- attribute uncertainty: modify Kriging system to account for (possibly varying) accuracy of each individual source datum (source = sum of points + error)
- support boundary uncertainty: stochastic sampling functions
Inference of Point Covariance

Top-down approach:
- iterative fitting of point covariance model, accounting for support differences
- difficult to get covariance model at short distances, unless point data exist

Bottom-up approach:
- spatial process theories (e.g., diffusion, fractals) dictate point covariance
- partial differential equations specify smoothness of solution surface given BCs

Hybrid approach:
- Fit different plausible theories to the data, and decide which one to keep based on some goodness-of-fit criteria
Interoperability and Uncertainty

In a perfectly accurate world, you hope that working with more data:

- will help you find more pertinent information
- hence, arrive at a better answer or decision

In an uncertain world, you hope that working with more non-conflicting data:

- will help you reduce uncertainty, and
- increase likelihood that confidence intervals bracket true answer

Spatial uncertainty is necessarily subjective (a.k.a. a model), and its impact is case-specific
Uncertainty in Spatial Interpolation

Unknown spatial distribution

Sample data

Uncertainty meta-surface

Interpolated spatial distribution
Looking at these two maps alone, no matter how well they are portrayed:

- does not give you the complete/correct picture of spatial patterns
- does not allow you to rigorously assess the impact of spatial uncertainty on decision-making, because local mean and variance are single pixel info.
Spatial Uncertainty Analysis (1)

- (single output) model with multiple inputs:
  \[ y = \phi(z_t) \]
  \( \phi(\cdot) = \text{arbitrary function or set of rules operating on many locations of an attribute map} \)

- vector of model inputs:
  \[ z_t = \left[ z\left(t_p\right), p = 1,\ldots,P \right]' \]
  pertaining to different locations or attributes, and of arbitrary support or quality

- data vector with observations:
  \[ z_s = \left[ z\left(s_k\right), k = 1,\ldots,K \right]' \]
  pertaining to different locations or attributes (possibly not those found in \( z_t \)), and of arbitrary support or quality; could include measurements of actual y-values
Spatial Uncertainty Analysis (2)

Uncertainty: if model inputs are random, so are the model outputs

Objective: characterize distribution of model outputs

Caveats: 
\[ E \{y\} \neq \phi \left( E \{z_t\} \right) \quad \text{Var} \{Y\} \neq \phi \left( \text{Var} \{z_t\} \right) \]

unless the model is linear even for linear models, you need covariances

- **Monte Carlo simulation** = generation of synthetic realizations (samples) from *joint* multivariate distribution of model inputs, given the sample data, and propagate each one of these realizations through the model:

\[
F_{1 \ldots P} \left( z_1, \ldots, z_P \mid z_s \right) = \text{Prob} \left\{ Z \left( t_1 \right) \leq z_1, \ldots, Z \left( t_P \right) \leq z_P \mid z_s \right\}
\]

Realizations should be as realistic as possible, i.e., reproduce known diverse data, histogram, and model of spatial correlation
Monte Carlo Uncertainty Propagation

- Generation of alternative, realistic images, consistent with all info available
- Numerical evaluation of model on each synthetic realization = *model response*
- Set of alternative model outputs = *model response distribution*

Alternative synthetic images

GIS operation, environmental model, engineering analysis

Data-constrained simulated realizations

Response distribution

distribution of alternative model outputs
Does Spatial Correlation Matter?

Misspecification of spatial correlation can lead to **biased** results (especially in non-linear models).
Uncertainty in Segmentation via Simple Thresholding (1)

Reference image

Sample data

Reference segmentation

Interpolated image

Cutoff = 0.85 quantile of distribution of reference values
Uncertainty in Segmentation via Simple Thresholding (2)
Uncertainty in Segmentation via Simple Thresholding (3)

=> Simulated images are realistic stimuli for segmentation by thresholding; Interpolated images are generally not (depending on the sample density)
Uncertainty in Segmentation via Simple Thresholding (4)

Local mean

Local variance

Segmentation

Probability > cutoff
Making Sense Out of Spatial Uncertainty Assessment

A two-step data mining endeavor

- Mining the inputs to the system (or realizations of RF model):
  Stochastic simulation for generating realistic input stimuli, consistent with spatial correlation and available data

- Mining the outputs of the system (model outputs):
  Analysis of objects resulting from image segmentation / pattern recognition, e.g., calculation of areas, complexity of perimeter, etc…
Decision-Making Under Uncertainty

Multiple stochastic actions

Risk Attitude

Loss-function

Multiple expected benefits/losses

For each possible action:

1) pretend that action is your decision
2) treat all other actions as possible truths
3) compute loss associated with this action, had each possible truth been realized
4) compute expected (average) loss

Optimal Decision

Action minimizing expected loss
The Expected Value of Including Uncertainty (EVIIU)

Multiple stochastic actions

Risk Attitude → Loss-function → Optimal Decision
(1) Action minimizing expected loss

EVIU

Difference between expected benefits from (1) and (2)

Single deterministic action

Risk Attitude → Loss-function → Optimal Decision
(2) Deterministic action minimizing loss
More Useful Risk Concepts

Expected value of *perfect* information:

- **Definition:** difference between expected benefits associated with optimal decisions made after and before acquiring additional information
- **Requisite:** additional information must have been already acquired (posterior or terminal analysis)

Expected value of *sample* information:

- **Definition:** difference between expected benefits associated with optimal decisions made after and before acquiring *simulated* additional information or, *how much are you willing to pay to acquire this new information*
- **Requisite:** link between existing and additional information; the latter need not be acquired (pre-posterior analysis)

*All these concepts require spatial uncertainty modeling, NOT predictions*
A stochastic framework accounting for scale differences in interoperability:

- **general**: can handle integrated measurements over arbitrary domains
- **simple**: utilizes standard geostatistical theory with minor modifications
- **comprehensive**: can handle alternative types of point covariance models
- **consistent**: guarantees reproduction of data at larger scales (mass preserving)
- **providing uncertainty assessment**: regarding target predictions, and hence enables risk-conscious decision making