Spatial Prediction using Interpolation (Kriging)

Approaches to spatial prediction

This is the **prediction** of the value of some variable at an **unsampled point**, based on the values at the **sampled points**.

This is often called **interpolation**, but strictly speaking:

- **Interpolation**: prediction is only for points that are **geographically inside** the (convex hull of the) sample set;
- **Extrapolation**: prediction outside this geographic area

(Note: same usage as in feature-space predictions)

A taxonomy of spatial prediction methods

Strata divide area to be mapped into 'homogeneous' **strata**; predict **within each stratum** from all samples in that stratum

Global predictors: use **all samples** to predict at **all points**; also called **regional** predictors;

Local predictors: use only **'nearby'** samples to predict at each point

Mixed predictors: some of structure is explained by strata or globally, some locally

Approaches to prediction: Global (Regional) Predictors

- Value of the variable depends on **relative geographic position** within a spatial field
  - * Example: thickness of a layer of volcanic ash over a buried soil
  - * Example: amount of quartz gravels in a soil derived from conglomerate residuum
  - Groundwater depth below a reference level
- Since there is only one process (global), **all** sample points are used to compute the prediction
- (Sometimes the sample points are limited to a region, e.g. in moving trend surfaces or splines)

Polynomial trend surfaces

- A global predictor which models a **regional trend**
- The value of a variable at each point depends only on its **coordinates** and parameters of a fitted surface
- This is modeled with a smooth function of position,
  $z = f(x,y) = f(\text{East},\text{North})$
  for grid coordinates; this is called the **trend surface**
- Simple form (plane, 1st order):
  $z = \beta_0 + \beta_x E + \beta_y N$
- Higher-order surfaces may also be fitted (beware of fitting the noise!)

Approaches to prediction: Local predictors

- **No strata**
- **No regional trend**
- Value of the variable is predicted from **"nearby" samples**
  - * Example: concentrations of soil constituents (e.g. salts, pollutants)
  - * Example: vegetation density

Local Predictors

Each interpolator has its own assumptions, i.e. theory of spatial variability

- Nearest neighbour (Thiessen polygons)
- Average within a radius
- Average of the n nearest neighbours
- Distance-weighted average within a radius
- Distance-weighted average of n nearest neighbours
- . . .
- "Optimal" weighting) **Kriging**

Nearest neighbor (Thiessen polygons)

- Predict each point from its **single nearest sample point**
- Conceptually-simple, makes the minimal assumptions about spatial structure
- No error estimate possible, ignores other 'nearby' information
- Maps show abrupt discontinuities at boundaries, so don't look very realistic
- But may be a more accurate predictor than poorly-modelled predictors

Average within a radius

- Use the set of all neighbouring sample points within some radius r
- Predict by averaging : $\hat{x_0} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad d(x_0, x_i) \le r$

Although we can calculate error estimates from the neighbors, these assume no spatial structure closer than the radius

- Problem: How do we select a radius?

Inverse Distance-weighted (IDW)

- Inverse of distance to some set of n nearest-neighbors:

$$\hat{x_0} = \sum_{i=1}^{n} \frac{x_i}{d(x_0, x_i)} \Big/ \sum_{i=1}^{n} \frac{1}{d(x_0, x_i)}$$

- Inverse of distance to some set of n nearest-neighbors, to some power k

$$\hat{x_0} = \sum_{i=1}^{n} \frac{x_i}{d(x_0, x_i)^k} \Big/ \sum_{i=1}^{k} \frac{1}{d(x_0, x_i)^k}$$

- Implicit theory of spatial structure (a **power model**), but this is **not testable**
- Can select all points within some limiting distance (radius), or some fixed number of nearest points, or . . .
- How to select radius or number and power?

Problems with above Methods

Problems with average-in-circle methods:

1. No objective way to select radius of circle or number of points

- Problems with inverse-distance methods:
  1. How to choose power (inverse, inverse squared . . .)?
  2. How to choose limiting radius?

- In both cases:
  1. Uneven distribution of samples could over– or under–emphasize some parts of the field
  2. prediction error must be estimated from a separate validation dataset

Ordinary Kriging (OK)

- The theory of regionalized variables leads to an "optimal" interpolation method, in the sense that the prediction variance is minimized.
- This is based on the **theory of random functions**, and requires certain **assumptions**.
- Dealing with following questions:
  1. In what sense is OK "optimal"?
  2. Derivation of the OK system of equations
  3. Interpolation by kriging

An "optimal" local predictor

- Prediction is made as a **linear** combination of known data values (a **weighted average**).
- Prediction is **unbiased** and **exact at known points**
- Points closer to the point to be predicted have larger weights
- Clusters of points "reduce to" single equivalent points, i.e., over-sampling in a small area can't bias result
- Closer sample points "mask" further ones in the same direction
- Error estimate is based only on the sample configuration, not the data values
- **Prediction error should be as small as possible**.

Kriging

- A **"Best Linear Unbiased Predictor"** (BLUP) that satisfies certain criteria for optimality.
- **It is only "optimal" with respect to the chosen model!**
- Based on the **theory of random processes**, with **covariances depending only on separation** (i.e. a variogram model)
- Theory developed several times (Kolmogorov 1930's, Wiener 1949) but current practice dates back to Matheron (1963), formalizing the practical work of the mining engineer D G **Krige** (RSA).

* Should really be written as "krigeing" (Fr. **krigeage**) but it's too late for that.

How do we use Kriging?

1. **Sample**, preferably at different resolutions
2. **Calculate** the **experimental variogram**
3. **Model** the variogram with one or more authorized functions
4. **Apply** the kriging system, with the variogram model of spatial dependence, at each point to be predicted
   - Predictions are often at each point on a **regular grid** (e.g. a raster map)
   - These 'points' are actually blocks the size of the sampling support
   - Can also predict in **blocks** larger than the original support
5. Calculate the **error** of each prediction; this is based only on the **sample point locations**, not their data values.

Prediction with Ordinary Kriging (OK)

In OK, we model the value of variable z at location s_i as the sum of a **regional mean** m and a **spatially-correlated random component** $e(s_i)$:

$$Z(s_i) = m + e(s_i)$$

- The regional mean m is estimated from the sample, but not as the simple average, because there is spatial dependence. It is **implicit** in the OK system.

Stationarity

- Restrictions on the nature of spatial variation that are required for OK (among others) to be correct
- **First-order**: the **expected values** (mean) at all locations in the field are the same:

$$E[Z(\vec{x}_i)] = \mu, \forall \vec{x}_i \in R$$

- **Second-order**:
  1. The **variance** at any point is finite and the same at all locations in the field
  2. The covariance structure depends only on **separation** between point pairs:
     - This makes the prediction model (previous slide) valid.

Ordinary Kriging (OK)

- Predict at points, with unknown mean (which must also be estimated) and no trend
- Each point $\vec{x}_0$ is predicted as the **weighted average** of the values at all sample points

$$\vec{x}_i: \hat{Z}(\vec{x}_0) = \sum_{i=1}^{N} \lambda_i z(\vec{x}_i)$$

The weights l assigned to each sample point sum to 1:

$$\sum_{i=1}^{N} \lambda_i = 1$$

- Therefore, the prediction is **unbiased**:

$$E[\hat{Z}(\vec{x}_0) - Z(\vec{x}_0)] = 0$$

- "Ordinary": no trend or strata; regional mean must be estimated from sample

Prediction variance

- Depends on the variogram function γ(h) and the point configuration **around each point** to be predicted:

$$\begin{aligned}\mathrm{Var}[\hat{Z}(\vec{x}_0)] &= E[\{\hat{Z}(\vec{x}_0) - Z(\vec{x}_0)\}^2] \\ &= 2\sum_{i=1}^{N}\lambda_i\gamma(\vec{x}_i,\vec{x}_0) - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\gamma(\vec{x}_i,\vec{x}_j)\end{aligned}$$

- First term: **lower semi-variances between a point and the sample points leads to a lower prediction variance**; different for each point to be predicted
- Second term: **respect the co-variance structure of the sample points**; depends on configuration of sample points only
- We do not yet know what are the optimal weights l, but once we do, we can calculate this prediction variance; so **we can selected the** l **to minimize it**.

'Model globally, predict locally'

- The **kriging equations** are solved **separately for each point** $\vec{x}_0$, using the semivariances around that point, in a local neighbourhood; this gives a **different** set of weights λ for each point to be predicted.
- However, the **variogram model** γ() used in these equations is estimated **only once**, using information about the spatial structure over the whole study area.
- Q: How is this possible?
  - * A1: Assume **second-order stationarity**
  - * A2: Assume at least **local first-order stationarity** (local weights will be high enough to mask long-distance non-stationarity)

Computing the weights

- There is one important piece of the puzzle missing: **How do we set the weights** l around a point to be predicted?
- Recall the ad-hoc method: some power of inverse distance
- We want these to be the "best", based on an **objective function** that mathematically defines what we mean by "best".
- There will be an optimum combination of weights at the point to be predicted, given the **point configuration** and the **modelled variogram**
- We compute these weights for each point to be predicted, by an **optimization criterion**, which in OK is **minimizing the prediction variance**.

Objective function (1): Unconstrained

- In a minimization problem, we must define an **objective function** to be minimized. In this case, it is the prediction variance in terms of the N weights λ_i:

$$\begin{aligned}f(\lambda) &= \mathrm{var}[\hat{Z}(\vec{x}_0)] \\ &= 2\sum_{i=1}^{N}\lambda_i\gamma(\vec{x}_i,\vec{x}_0) - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\gamma(\vec{x}_i,\vec{x}_j)\end{aligned}$$

- But this is unbounded and can be trivially solved by setting all weights to 0. We must add another constraint to bound it.

Objective function (2): Constrained

To bound the objective function, we need another constraint; here it is naturally **unbiasedness**.

This is added to the system with a **LaGrange multipler** ψ:

$$\begin{aligned}f(\lambda,\psi) &= \mathrm{var}[\hat{Z}(\vec{x}_0)] - 2\psi\left\{\sum_{i=1}^{N}\lambda_i - 1\right\} \\ &= 2\sum_{i=1}^{N}\lambda_i\gamma(\vec{x}_i,\vec{x}_0) - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\gamma(\vec{x}_i,\vec{x}_j) - 2\psi\left\{\sum_{i=1}^{N}\lambda_i - 1\right\}\end{aligned}$$

Note that the last term = 0, i.e. the prediction is unbiased.

Minimization

This is now a system with N +1 unknowns.

Minimize by setting all N +1 partial derivatives to zero:

$$\begin{aligned}\frac{\partial f(\lambda_i,\psi)}{\partial\lambda_i} &= 0, \forall i \\ \frac{\partial f(\lambda_i,\psi)}{\partial\psi} &= 0\end{aligned}$$

In the last differential equation, all the λ are constants, so the first two terms differentiate to 0; in the last term the ψ differentiates to 1 and we are left with the unbiasedness condition:

$$\sum_{i=1}^{N}\lambda_i = 1$$

The Kriging system

In addition to unbiasedness, the partial derivatives give N equations (one for each λ_i) in N +1 unknowns (the λ_i plus the LaGrange multiplier ψ):

$$\sum_{j=1}^{N} \lambda_j \gamma(\vec{x}_i,\vec{x}_j) + \psi = \gamma(\vec{x}_i,\vec{x}_0), \quad \forall i$$

This is now a system of N +1 equations in N +1 unknowns and can be solved by standard linear algebra.

The **semivariances between sample points** $\gamma(\vec{x}_i, \vec{x}_j)$ are computed **only once** for any point configuration; however the **semivariances at a sample point** $\gamma(\vec{x}_i, \vec{x}_0)$ must be **computed separately for each point to be predicted**.

Solving the Kriging system

At each point to be predicted:

1. **Compute the semivariances** γ from the separation between the point and the samples, according to the **modelled variogram**
2. **Solve simultaneously** for the weights and multiplier
3. **Compute the predicted value** as the **weighted average** of the samples
4. **Compute the variable term** of the prediction variance
5. **Add the constant term** of the prediction variance to get the total variance.

Importance of the variogram model

- The kriging system is solved using the modeled semi-variances
- **Different models will give different kriging weights** to the sample points . . .
- . . . and these will give different predictions
- Conclusion: **bad model** leads to **bad predictions**

Matrix form of the Ordinary Kriging system

$$\mathbf{A}\lambda = \mathbf{b}$$

$$\mathbf{A} = \begin{bmatrix} \gamma(\vec{x}_1,\vec{x}_1) & \gamma(\vec{x}_1,\vec{x}_2) & \cdots & \gamma(\vec{x}_1,\vec{x}_N) & 1 \\ \gamma(\vec{x}_2,\vec{x}_1) & \gamma(\vec{x}_2,\vec{x}_2) & \cdots & \gamma(\vec{x}_2,\vec{x}_N) & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \gamma(\vec{x}_N,\vec{x}_1) & \gamma(\vec{x}_N,\vec{x}_2) & \cdots & \gamma(\vec{x}_N,\vec{x}_N) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \psi \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} \gamma(\vec{x}_1,\vec{x}_0) \\ \gamma(\vec{x}_2,\vec{x}_0) \\ \vdots \\ \gamma(\vec{x}_N,\vec{x}_0) \\ 1 \end{bmatrix}$$

How realistic are maps made by Ordinary Kriging?

- The resulting surface is **smooth** and shows **no noise**, no matter if there is a nugget effect in the variogram model
- So the field is the **best at each point taken separately**, but taken as a whole is **not a realistic map**
- The **sample points are predicted exactly**; they are assumed to be without error, again even if there is a nugget effect in the variogram model

* Note: block kriging does not have this problem

Topic: Block Kriging

- Often we want to predict in blocks of some defined size, not at points. Block kriging (BK) is quite similar in form to OK, but the estimation variances are lower.
- Estimate at blocks of a defined size, with unknown mean (which must also be estimated) and no trend
- Each **block** B is estimated as the weighted average of the values at all sample **points** x_i:
- As with OK, the weights λ_i sum to 1, so that the estimator is unbiased, as for OK

Mixed interpolators

There is quite some controversy about the use of the following terms, and you may well find them used in a different way than the following.

- **Universal Kriging** (UK): includes a global trend as a function of the geographic coordinates. (Note: Some authors, including gstat, use this term for all mixed methods.)
- **Kriging with External Drift** (KED): includes feature-space predictors that are not geographic coordinates. (Note: Some authors use this term for all mixed methods, and consider UK a special case, where the predictors are coordinates.)
- **Regression Kriging** (RK), also called "kriging after de-trending" models the **trend** (geographic or feature space) and its **residuals** separately.

Simple Kriging

In OK we must estimate the regional mean along with the predicted values, in one OK system.

In UK or KED we must estimate both the intercept (b0) and all other trend coeficients (bi), along with the predicted values, in one UK system.

However, there may be situations where **the regional mean is known**. Then we can use so-called **Simple Kriging** (SK)

Similarly, if **the trend is known**, we can use "Simple" variants of UK and KED.

Exercises

Computer program – GS + 8