Applied geostatistics

Lecture 8 – Spatial sampling

D G Rossiter
University of Twente.
Faculty of Geo-information Science & Earth Observation (ITC)

October 16, 2010

Copyright © 2007–2010 University of Twente, Faculty ITC.
All rights reserved. Reproduction and dissemination of the work as a whole (not parts) freely permitted if this original copyright notice is included. Sale or placement on a web site where payment must be made to access this document is strictly prohibited. To adapt or translate please contact the author (`http://www.itc.nl/personal/rossiter`).

Topics for this lecture

1. Design-based sampling in the presence of spatial dependence
2. Optimal sampling grid with known variogram
3. *Spatial Simulated Annealing*
4. *Sampling for mixed models of spatial dependence*
5. *Sampling to model the variogram*

The topics written *in italic script* are supplementary; there is not enough time to cover them in one lecture. They belong logically here, so are included for your future reference.

Spatial sampling

A sample should be designed to extract the **maximum information** about reality from a **small portion** of it, with a miniumum of cost and effort.

Optimal **non-spatial** sampling has a long history; this must be extended to the case where there is **spatial dependence between observations**.

Commentary

Our knowledge of nature comes from **samples**, that is, a set of **sampling units** (sometimes loosely called "samples") taken from the **sampling frame**, which is all possible units in the underlying **population**. On the basis of the sample we make **inferences** about the population; e.g. we predict on a grid of many thousands of locations based on a sample of a few hundred observations.

Sampling is expensive and time-consuming, so designing a good sampling scheme (maximum information at minimum cost) is an important application of geostatistics

References

- Texts
  - de Gruijter, J.; Brus, D.J.; Bierkens, M.F.P. and Knotters, M., 2006. *Sampling for Natural Resource Monitoring*. Springer.
  - Cochran, W. G. 1977. *Sampling Techniques* (3rd ed.). New York: John Wiley.
  - Schreuder, H. T; Ernst, R and Ramirez-Maldonado, H., 2004. *Statistical techniques for sampling and monitoring natural resources*. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station; `http://www.fs.fed.us/rm`

- Journal articles

  - Minasny, B., McBratney, A.B. and Walvoort, D.J.J., *The variance quadtree algorithm: Use for spatial sampling design*. Computers & Geosciences, In Press, Corrected Proof.
  - Minasny, B. and McBratney, A.B., 2006. *A conditioned Latin hypercube method for sampling in the presence of ancillary information*. Computers & Geosciences, 32(9): 1378-1388.
  - Brus, D.J. and de Gruijter, J.J., 1997. *Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil* (with Discussion). Geoderma, 80(1-2): 1-59.
  - Stein, A. and Ettema, C., 2003. *An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons*. Agriculture, Ecosystems & Environment, 94(1): 31-47.

Why sample?

1. To make some statement about the **area as a whole**
   - Average, total, or variance; e.g. total biomass; average biomass per ha
2. To **map** some variable over an area
   - Predicted values at points: mean, confidence intervals, probability of exceeding a threshold ...
   - Same, for **block averages**
3. To determine **spatial structure**: direction, range, strength of dependence → causes, processes

Different statements have **different sampling requirements**

To check your understanding . . .

Q1 : *What is the main purpose of sampling in an area that has never been sampled before? What important property should the sample have?* *Jump to A1* •

Q2 : *What is the main purpose of additional sampling in a heavily-sampled area? What important property should the sample have?* *Jump to A2* •

Commentary

Sampling depends on the **model of spatial variation** that we assume or for which we have evidence.

The classical Discrete model (DMSV) assumes no spatial structure, instead using feature-space to design the sample; the Continuous model (CMSV) only recognizes spatial structure; and the Mixed model (MMSV) includes both.

We now look at each of these in detail.

Models of spatial variation

- **Discrete model** (DMSV): crisp boundaries, discrete classes, no spatial dependence within polygons; variance within class estimated by all samples from the class
  - **"Design-based"** sampling, based on feature-space structure (e.g. strata, continuous feature-space predictors)
- **Continuous model** (CMSV): no boundaries, no classes, spatial dependence to some range, all spatial variability is found by the variable itself. May include a global (trend) and local component
  - **"Model-based"** sampling ('model' of spatial dependence)
- **Mixed** model (MMSV)
  1. Stratify by DMSV, model within by CMSV
  2. Design-based methods incorporating spatial structure

What is different about designs considering geostatistics?

The spatial structure is modelled, so . . .

1. DMSV: For **design-based** methods we can avoid spatial dependence, so that our estimates of error are not biased

2. CMSV: We can place samples for maximum information or minimum cost in a **model-based** (geostatistical) sample

3. MMSV: Must consider both spatial dependence in **geographic space** and the spread of samples in **feature space**

Spatial sampling: DMSV

Here we assume all locations in a stratum are the same; therefore there is no spatial dependence.

Sampling designs with the DMSV

This is classical statistics applied to points that are located in geographic space, but where this is not considered a predictive factor, i.e. **no spatial dependence**

- Ignore spatial distribution; consider only **feature space**
- May consider predictive factors (design) in **feature space**; leads to **stratified** sampling designs
- Can use "classical" (non-spatial) inference, e.g. multiple regression in feature space

However, there is often **evidence** that the **residuals** (after the stratification or feature-space prediction) are **spatially-correlated**; then a scheme based on the MMSV should give more **efficiency** (see below).

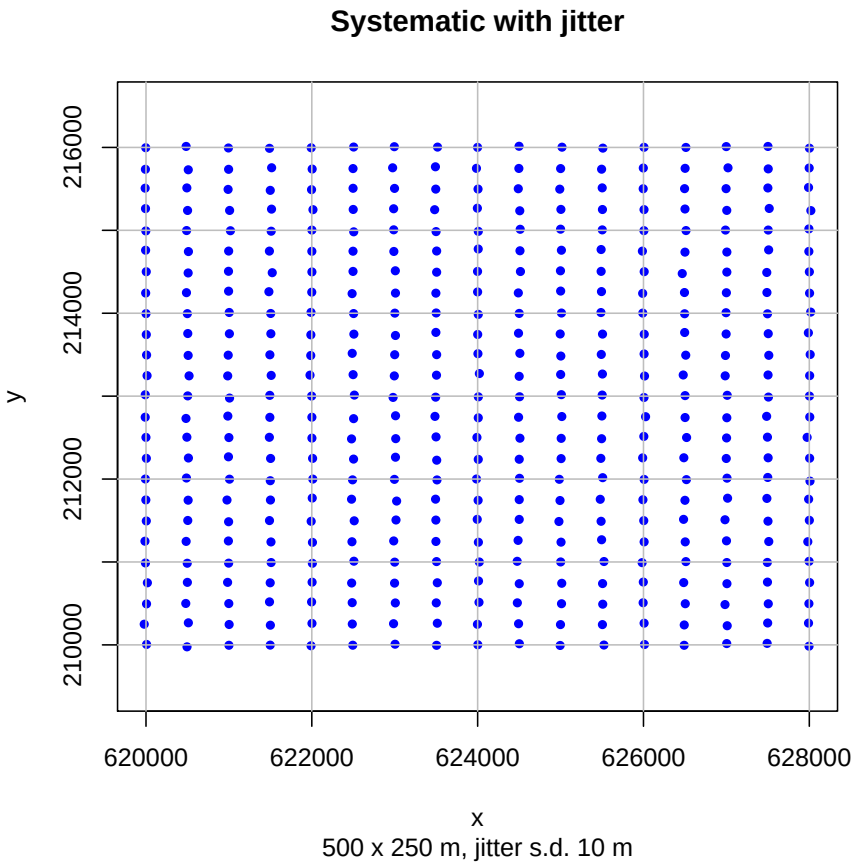
Unstratified sampling

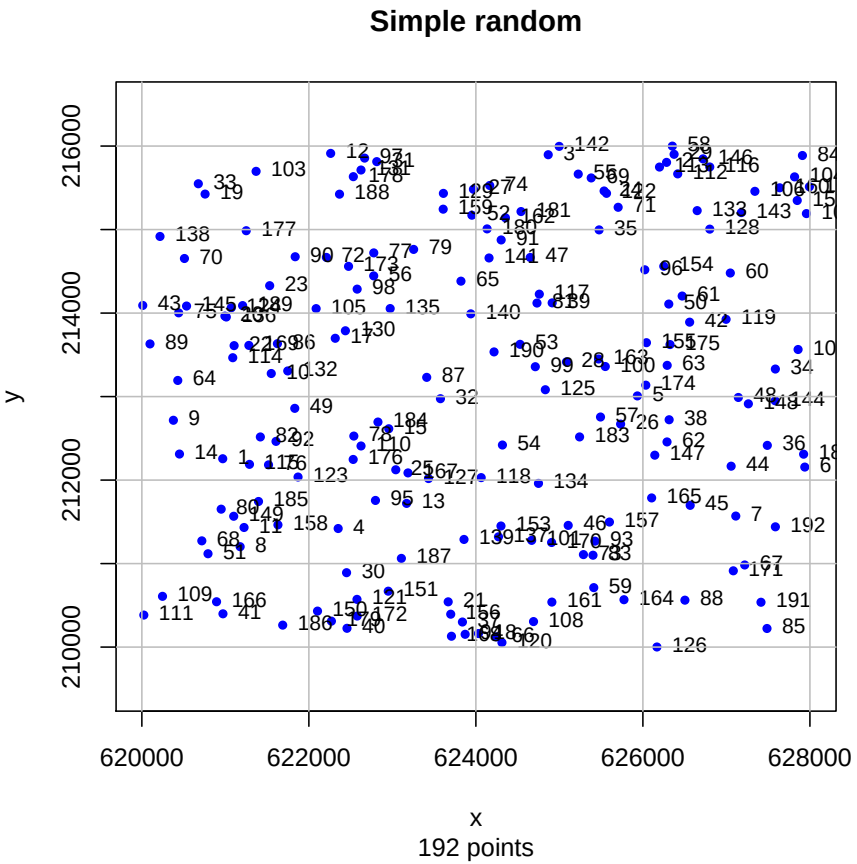
- Used when there are no sub-areas or feature-space classifiers

- Three main design classes:

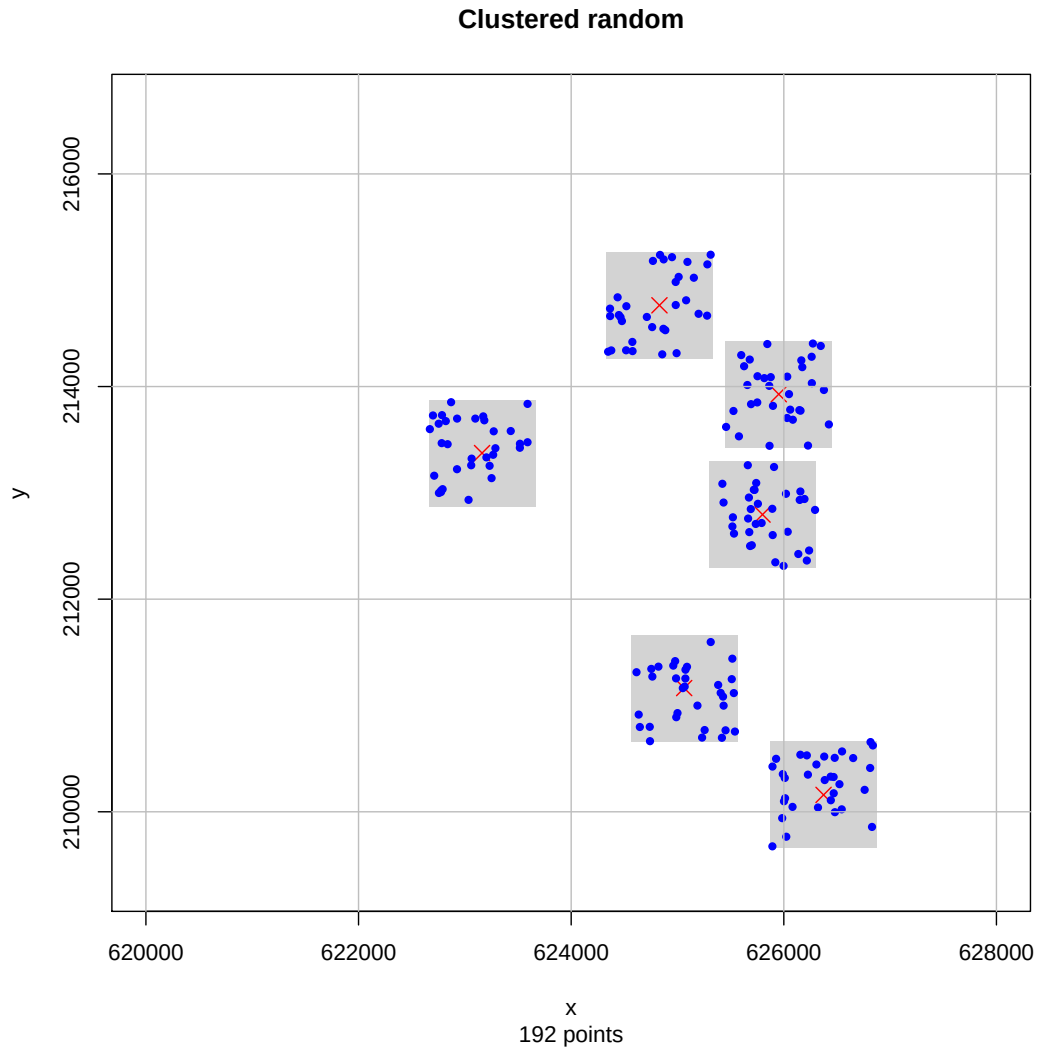
1. **Completely random**
   - has observations at different spatial ranges
   - may leave "holes" so not optimal for **mapping**
   - gives unbiased estimates of **population** statistics, e.g. mean
2. **Regular grid**
   - covers the area evenly, suited for mapping; can not estimate spatial structure
   - can add a small random **jitter** to each location, to avoid undected linear features
3. **Clustered**
   - either centres of clusters or samples near cluster centres can be random; the other is typically a grid.
   - Advantages: easier logistics; covers whole area but has observations at different spatial ranges

Two unstratified sampling designs





Example of a clustered spatial design



Six centres; random within 1 km^2 around random centre; 192 total observations

Stratified sampling

Often we have prior evidence that the whole space **should not be considered as a homogeneous unit** for sampling:

- There are **sub-areas** where there is evidence that the target variables should be different
  - E.g. in a soil survey the underlying lithology as shown on a geologic map is expected to influence the soil texture; the land-use history as shown on a land-use map is expected to influence the soil organic matter;
- There are **feature-space** attributes that are expected to influence the target variables
  - E.g. in a soil survey, the terrain slope gradient is expected to influence the soil depth (steeper slopes → shallower soils)

How to divide the sample among strata?

If we have a certain number of samples we can take, how to decide how many to allocated to each stratum?

1. **Proportional** to the area covered by the stratum: **proportional** stratified sampling; then the same formulas as for unstratified sampling can be applied;

2. **Proportional to the variance** within a stratum: more samples to more variable strata. This requires an *a priori* estimate of within-stratum variance. It concentrates the sampling effort where more samples are needed to characterize.

3. **According to interest** in the stratum. E.g. soils in agricultural areas will be more intensively used than in conservation areas, so more samples might be selected in the ag. areas.

Avoiding spatial dependence in DMSV sampling designs

1. Model variogram to find *range*: distance at which there is no dependence

2. Distribute points by a design-based method, e.g. Completely Random, Stratified Random, Area-Frame ...

3. **Restriction**: reject any point that is within the range of a point already selected

Design-based estimates of the error, from classical sampling theory, will be correct; otherwise they would be too favourable.

Spatial sampling for the CMSV

Here we assume that the only structure is **spatial**

That is, the target attribute is the result of a **random field**.

Sampling designs with the CMSV: objectives

1. Maximize information
   - Cover the largest possible area
   - Minimize some **optimization criterion**
2. Minimize costs
3. Incorporate any existing sample (avoid duplication)

What is to be optimized?

An **optimization criterion** is some numerical measure of the quality of the sampling design. Some possibilities:

1. Minimize the **maximum kriging variance** in the area: nowhere is more poorly predicted than this maximum

2. Minimize the **average kriging variance** over the entire area

3. Maximize the **information in a sample variogram**, to allow reliable variogram estimation.

Optimal point configuration (CMSV)

In a square area to be mapped, given a fixed number of points that can be sampled, in the case of bounded spatial dependence:

- Points should in on some **regular pattern**; otherwise some points duplicate information at others (in kriging, will "share" weights)
- Optimal (for both the "minimal maximum" and "minimal average" criteria): **equilateral triangles** (If the triangle is 1^2, max. distance to a point $= \sqrt{7}/4 \approx 0.661$)
- Sub-optimal but close: **square grid** (max. distance $= \sqrt{2}/2 \approx 0.707$)
  - Grid should be slightly **perturbed** so samples do not line up exactly; avoids unexpected periodic effects

(Problems: edge effects in small areas; irregular areas.)

Optimal point configuration in the presence of anisotropy

Optimal designs are easily adjusted for anisotropy.

The regular grid may be adapted for affine or **geometric anisotropy**: stretch it in the direction of maximum dependence, based on the **anisotropy ratio**.

E.g. for a ratio of 0.5, squares become rectangles, with the distance in the direction with the longest range twice that of the shortest range.

Computing an optimal grid size (model-based design)

- Reference: **McBratney, A. B. & Webster, R.** (1981) "The design of optimal sampling schemes for local estimation and mapping of regionalized variables - I and II". *Computers and Geosciences*, **7**(4), 331-334 and 335-365; also in Webster & Oliver.

- In kriging, the estimation error is based **only** on the sample **configuration** and the chosen **model** of spatial dependence, not the actual data values

- So, **if** we know the spatial structure (variogram model), we can compute the maximum or average kriging variances **before** sampling, i.e. before we know any data values.

- Then we can make sampling decisions on the basis of **cost-benefit**

Error variance

- Recall: The kriging variance at a **point** is given by:

$$\begin{aligned}\hat{\sigma}^2(\vec{x}_0) &= \mathbf{b}^T\lambda \\ &= 2\sum_{i=1}^{N}\lambda_i\gamma(\vec{x}_i,\vec{x}_0) - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j\gamma(\vec{x}_i,\vec{x}_j)\end{aligned}$$

- This depends only on the **sample distribution** (what we want to optimise) and the **spatial structure** (modelled by the semivariogram)

- In a **block** this will be lowered by the **within-block** variance $\overline{\gamma}(B,B)$

Reducing kriging error

Once a regular sampling scheme is decided upon (triangles, rectangles, . . .), the kriging variance is decreased in two ways:

1. **reduce the spacing** (finer grid) to reduce semivariances; or

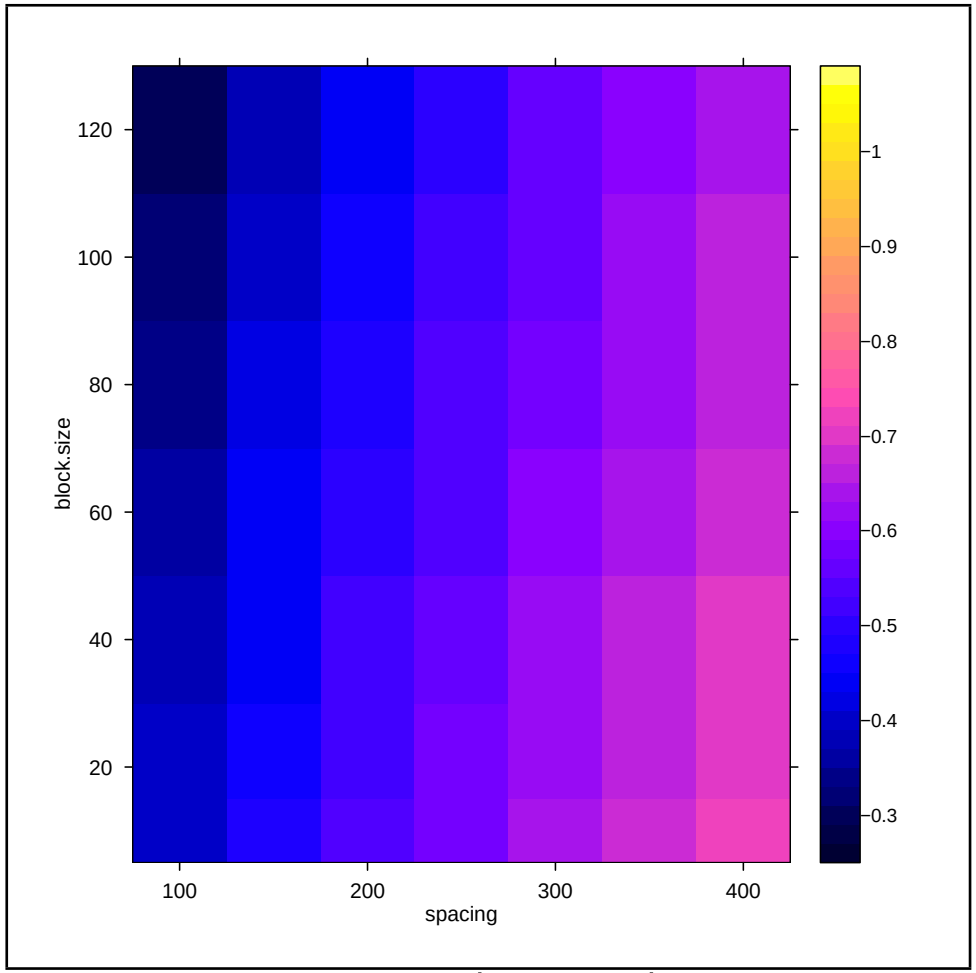
2. **increase the block size** of the prediction

These can be traded off; but usually the largest possible block size is selected, based on the mimimum decision area.

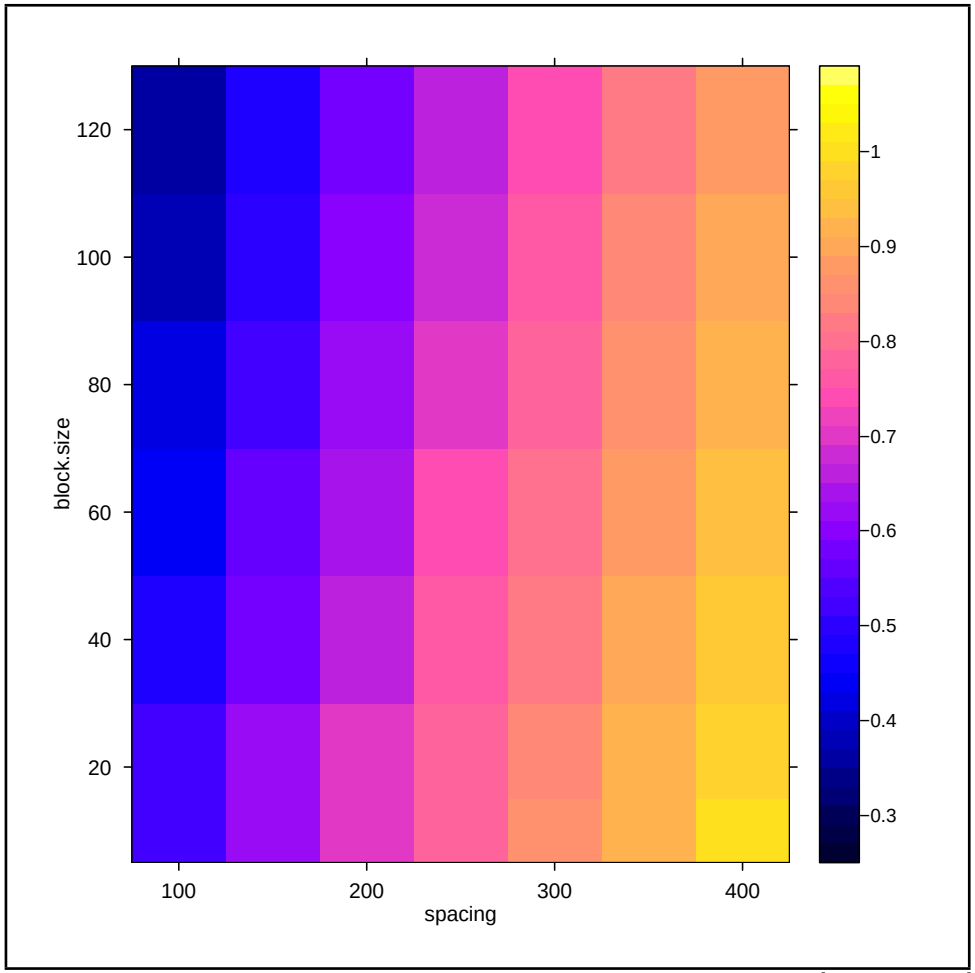
Error as a function of increasing grid resolution

- Consider 4 sample points in a square
- To estimate is one prediction point in the middle (furthest from samples → highest kriging variance)
- Criterion is "minimize the maximum prediction error"
- If the variogram is close-range, high nugget, low sill, we need a **fine** grid to take advantage of spatial dependence; high cost
- If the variogram is long-range, low nugget, high sill, a **coarse** grid will give similar results

Kriging variances at centre point



long range variogram (1200 m)



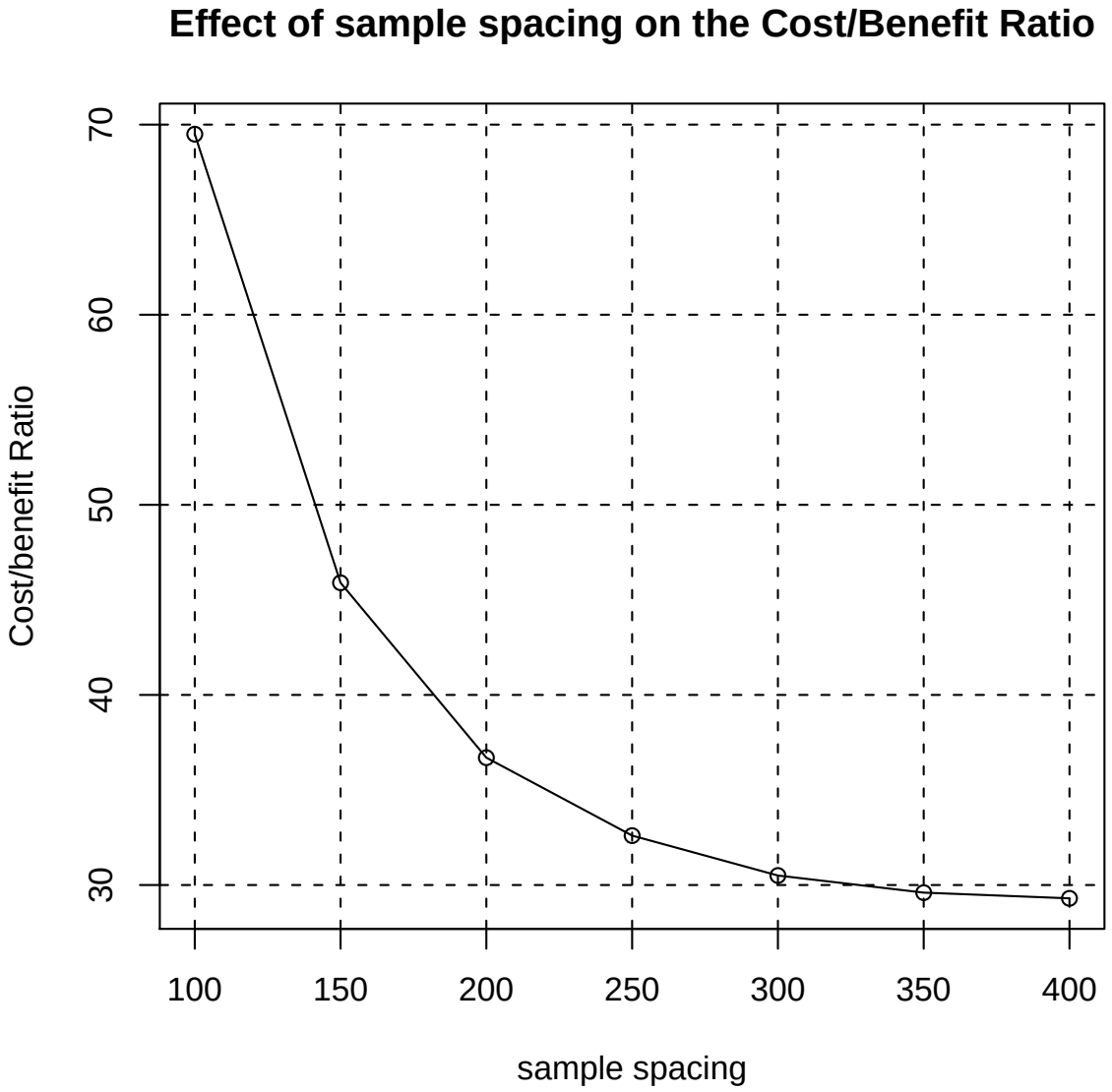
short range variogram (600 m)

Cost of mapping an area

- Given sample spacing (side of grid) g and total area A, the number of sample points required to cover the area is $n = (\frac{\sqrt{A}}{g} + 1)^2$
- Example: 25x25km area ($A = 625\text{km}^2$)
  - $g = 5\text{km} \rightarrow ((25/5) + 1)^2 = 36$
  - $g = 0.5\text{km} \rightarrow ((25/0.5) + 1)^2 = 2601$
  - $g = 10\text{km} \rightarrow ((25/10) + 1)^2 = 12.25 \approx 12$
- Multiply this by the cost of each sample
  - Fixed per sample: time to acquire, equipment rental for this time, laboratory
  - Variable: travel time between samples
- In addition, there is a fixed cost to set up the sampling scheme

Cost-benefit analysis

- Compute a **cost/benefit ratio** and plot against a controllable parameter:
  - sample spacing at a given block size
  - block size for a given sample spacing



(Note: this depends on the variogram)

Exercise

At this point you should do **Exercise 8: Spatial sampling** which is provided on the module CD:

1. Design completely random, stratified random, and grid sampling schemes;

2. Find the "optimal" regular grid sample for a given variogram model;

As in all exercises there are **Tasks**, followed by R code on how to complete the task, then some **Questions** to test your understanding, and at the end of each section the **Answers**. Make sure you understand all of these.

Spatial Simulated Annealing (SSA)

Problem: how to **optimally** place a **limited number** of observations in a study area in order to extract the **maximum information** at **minimum cost**.

We consider here the information to be a map over some study area, made by **ordinary kriging** from the sample points; so the assumptions of the CMSV must be met.

Reference:

van Groenigen, J.-W. (2000). The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma*, 97(3-4), 223-236.

also contained in the PhD thesis:

van Groenigen, J.-W. *Constrained optimisation of spatial sampling* Enschede, NL: ITC.

Commentary

The approach can be extended for the mixed model (MMSV), see:

Brus, D. J., & Heuvelink, G. B. M. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138(1-2), 86–95

Problems with the "optimal" grid

The "optimal" grid presented in the previous section is optimal only in restricted circumstances. There are many reasons that approach might not apply:

- **Edge** effects: study area is not infinite
- **Irregularly-shaped** areas, e.g. a flood plain along a river
- **Off-limits** or **uninteristing** areas, e.g. in a soils study: buildings, rock outcrops, ditches ...
- **Existing samples**, maybe from a preliminary survey; don't duplicate the effort!

Impossible to compute an optimum analytically (as for the regular grid on an infinite plane).

Annealing

Slowly cooling a molten mixture of metals into a stable crystal structure.

During annealing the **temperature** is slowly lowered.

At **high** temperatures, molecules move around rapidly and long distances

At **low** temperatures the system stabilizes.

Critical factor: speed with which temperature is lowered

- too fast: stabilize in a sub-optimal configuration
- too slow: waste of time

Simulated annealing

This is a numerical analogy to actual annealing:

- Some aspect of a numerical system is perturbed
- The configuration should approach an optimum
- The amount of perturbation is controlled by a "temperature"

Outline of SSA

1. Decide on an **optimality** criterion

2. Place the desired number of sample points "anywhere" in the study area (grid, random ...); compute **fitness** according to optimality criterion

3. Repeat (**iterate**):

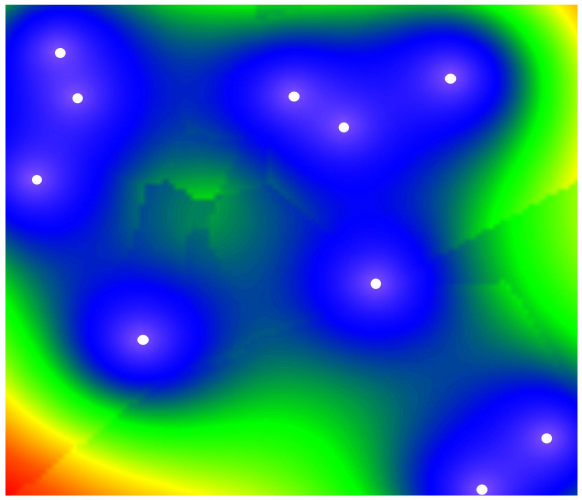
   (a) Select a point to move; **move** it a **random distance and direction**
   (b) If outside study area, try again
   (c) Compute **new fitness**
   (d) If **better**, accept new plan; if **worse** also accept with a certain **probability**

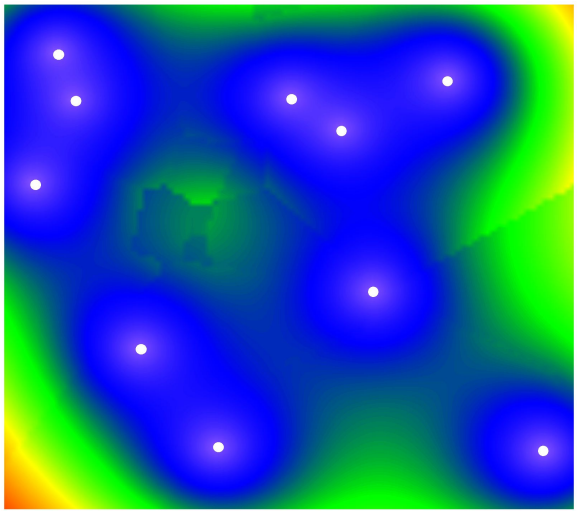
4. **Stop** according to some **stopping criterion**

Example of a single step

Colour ramp is from blue (low kriging variance) to red (high).

Point at lower right is moved to middle-bottom:





A large "hot" area (high kriging variance) is now "cooler".

Temperature

The distance to move a point is controlled by the **temperature**; this is used to multiply some distance.

$$T_{k+1} = \alpha \cdot T_k \tag{1}$$

where k is the step number and $\alpha < 1$ is an empirical factor that reduces the temperature; we must also specify an initial temperature T_0.

Fitness

Several choices, all based on the **kriging variance**:

- **Mean** over the study area (MEAN_OK)
  - appropriate when estimating spatial averages to a given precision
- **Maximum** anywhere in the study area (MAX_OK)
  - appropriate when the entire area must be mapped to a given precision, e.g. to guarantee there is no health risk in a polluted area.

Stopping criterion

Possiblities:

- fixed number of iterations
- reach a certain (low) temperature
- after a certain number of iterations with no change.

Acceptance criterion

Metropolis criterion: the probability $P(S_0 \rightarrow S_1)$ of accepting the new scheme is:

$$\begin{aligned} P(S_0 \rightarrow S_1) &= 1, \text{ if } \phi(S_1) \leq \phi(S_0) \\ P(S_0 \rightarrow S_1) &= \exp\left(\frac{\phi(S_0) - \phi(S_1)}{c}\right), \text{ if } \phi(S_1) > \phi(S_0) \end{aligned} \quad (2)$$

where S_0 is the fitness of the current scheme, S_1 is the fitness of the proposed new scheme, and c is the **temperature**. This can also be written:

$$p = e^{-\Delta f / T_k} \quad (3)$$

where T_k is the current temperature and Δf is the change in fitness due to the proposed new scheme.

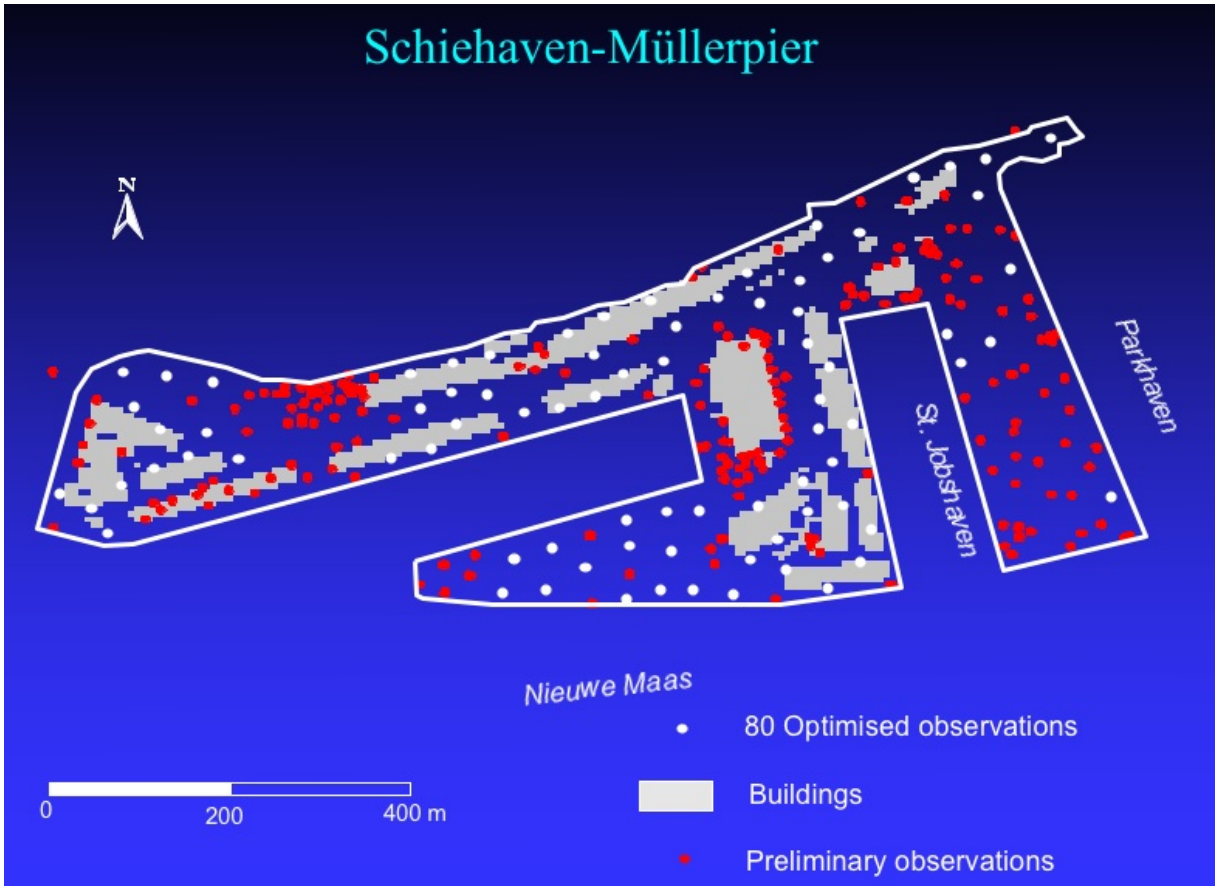
Note that this will be positive for a poorer solution, so its complement is used for the exponent.

To check your understanding . . .

Q3 : *Why should the procedure ever accept a* **worse** *scheme?* *Jump to A3* •

A real example

Industrial area, existing samples; more must be taken to lower the prediction variance to a target level everywhere; **where** to place the new samples?



Reference: **van Groenigen, J. W., Stein, A., & Zuurbier, R.** (1997). Optimization of environmental sampling using interactive GIS. *Soil Technology*, 10(2), 83-97

Sampling for the Mixed Model of Spatial Variability (MMSV)

The MMSV has both feature and geographic space dependence, so both must be considered when setting up a spatial sampling scheme.

Motivation for MMSV designs

Regression assumes that the errors from the model are *statistically independent*; this is often not be plausible, due to spatial dependence in the sources of error, as demonstrated by the residual variogram.

Thus the errors are *too optimistic*, if any points are within the range of spatial dependence; i.e. there is less error than if the points were uncorrelated.

Recent references:

- **Brus, D. J., & Heuvelink, G. B. M.** Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138(1-2), 86–95
- **Lark, R. M.** (2000). Regression analysis with spatially autocorrelated error: simulation studies and application to mapping of soil organic matter. *International Journal of Geographical Information Science*, 14(3), 247–264.

Sampling for MMSV

For a feature-space model, we may want to place points at particular locations in the feature space, but these may be spatially-correlated.

Reasons:

- **cost**: blocks or transects are easy to set up and cheap to sample
- specific **areas of interest**: feature-space characteristics we want to sample may be only in small areas (e.g. hilltops).

Solution 1: If range is fairly small relative to transect size, separate observations by at least the range.

Solution 2: Use GLS to estimate the trend surface and then model the residuals.

Sampling to model the variogram

If an area has never been sampled for a variable of interest, we need to determine the **range of spatial dependence** in order to set up an efficient sampling scheme.

Note that such a sampling scheme is *not* intended to map an area; thus there may be large "holes" in the coverage; the interest is instead on determining spatial structure.

If a map is wanted, the variogram derived from this first sampling exercise can be used to design an optimum grid sample, as explained above.

Some recent references

- **Webster, R., Welham, S. J., Potts, J. M., & Oliver, M. A.** (2006). Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. *Computers & Geosciences*, 32(9), 1320-1333.

- **Lark, R. M.** (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma*, 105(1-2), 49-80.

- **van Groenigen, J.-W.** (1999). Sampling strategies for effective variogram estimation. In J.-W. van Groenigen (Ed.), *Constrained optimisation of spatial sampling* (pp. 105-124). Enschede, NL: ITC.

- **Müller, W. G., & Zimmerman, D. L.** (1999). Optimal designs for variogram estimation. *Environmetrics*, 10(1), 23-37.

- **Russo, D.** (1984). Design of an optimal sampling network for estimating the variogram. *Soil Science Society of America Journal*, 48(4), 708-716.

Nested spatial sampling

An efficient way estimate a variogram is with the following **nested** spatial sampling scheme. It is based on work from 1937, re-discovered and extended in 1990.

References

- Original work: Youden, W. J. & Mehlich A. (1937) *Selection of efficient methods for soil sampling*, Contributions of the Boyce Thompson Institute for Plant Research *9*: 59-70

- Summarized in R. Webster & M. Oliver (1990) *Statistical methods in soil and land resource survey*; Oxford University Press, Ch. 13

- Also in R. Webster and M. Oliver (2001) *Geostatistics for environmental scientists*, §5.3

- **Recent paper** re-stating the method: Webster, R. Welham, S. J., Potts, J. M.,& Oliver, M. A. (2006) *Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood*, Computers & Geosciences *32*: 1320-1333

Sampling scheme to establish spatial dependence

- Purpose: establish the structure of spatial dependence (e.g. variogram) ...
- ...with a minimal number of samples
- Not intended to map an area
- Instead, to discover the optimal spacing to map

How to design the nested sample

- Widest spacing s_1 is the 'station', which are assumed so far away from each other as to be spatially independent
  - furthest expected dependence ...
  - ... based on the landscape ...
  - ... and expected range of process to be modelled
- Closest spacing s_n is the shortest distance whose dependence we want to know

Geometric series

- A geometric series increases terms by multiplication
- It allows us to cover a wide range of distances (possible ranges) with a few stages.
- Increase spacing in *geometric series*:
  $s = \sqrt{s_1 \cdot s_n}$
- Fill in series with further geometric means

Geometric series: example

- First series: $s_1 = 600\text{m}$ (stations), $s_5 = 6\text{m}$ (closest)
- Intermediate spacing: $s_3 = \sqrt{6\text{m} \cdot 600\text{m}} = 60\text{m}$
- Series now $\{600\text{m}, 60\text{m}, 6\text{m}\}$
- Fill in with the geometric means
  - $s_2 = \sqrt{600\text{m} \cdot 60\text{m}} \approx 190\text{m}$
  - $s_4 = \sqrt{60\text{m} \cdot 6\text{m}} \approx 19\text{m}$
- Final series $\{600\text{m}, 190\text{m}, 60\text{m}, 19\text{m}, 6\text{m}\}$

Locating the sample points

- Objective: cover the landscape, while avoiding systematic or periodic features
- Method: random bearings from centres at each stage
- Stations can be along a transect if desired (no spatial dependence)
- From a centre at stage i (E_i, N_i), to find a point (E_{i+1}, N_{i+1}) at the next spacing s_{i+1}:
  - $\theta = \text{random_uniform}[0 \ldots 2\pi]$
  - $E_{i+1} = E_i + (s_{i+1} * \sin\theta)$
  - $N_{i+1} = N_i + (s_{i+1} * \cos\theta)$

Number of sample points

- Number of stations selected to cover the area of interest
- At each stage S_i, the next stage S_{i+1} has in principle **double** the samples
- One is for all the previous centres from stage $S_1 \ldots S_{i-1}$ and one is for the new centre from stage S_i
- So the total number doubles: half old, half new centres

Unbalanced sampling

- After the first 4 stages, use an unbalanced design
- Only half the centres at S_i $(i \geq 4)$ are further sampled at S_{i+1}
- This still covers the area, but only uses half the samples at the shortest ranges
- Number of pairs is still enough estimate short-range dependence

Number of sample points: example

- Five stages {600m, 190m, 60m, 19m, 6m}
- Nine stations: $n_1 = 9$
- Double at stages 2 . . . 4: $n_2 = 18, n_3 = 36, n_4 = 72$
- At stage 5, only use half the 72 centres, i.e. 36
- Total at stage 5: $72 + 36 = 108$ (would have been 144 with balanced sampling)

Numbering of sample points

- Use a five-digit number, one place for each distance:
  $\{1 - n_1\}\{0 - 1\}\{0 - 1\}\{0 - 1\}\{0 - 1\}$

- except no combination $\{1 - n_1\}0\{0 - 1\}\{0 - 1\}$ i.e. only half the stage 4 centres are used for stage 5

- Angles are measured from a point with n 1's to the point with $n + 1$ 1's, differing in exactly one place. I.e., one 0 becomes a 1.

- Example:

  10000 ⇒ 11000 (190m)
  10000 ⇒ 10100 (60m)
  10000 ⇒ 10010 (19m)
  11000 ⇒ 11001 (6m)

Nested ANOVA : Partition Variability by sampling level

- Linear model:

$$z_{ijk\ldots m} = \mu + A_i + B_{ij} + C_{ijk} + \cdots + Q_{ijk\ldots m} + \varepsilon_{ijk\ldots m}$$

- Link with regional variable theory (semivariances): m stages; d_1 shortest distance at mth stage; d_m largest distance at first stage

$$\begin{aligned} \sigma^2_m &= \gamma(d_1) \\ \sigma^2_{m-1} + \sigma^2_m &= \gamma(d_2) \\ &\vdots \\ \sigma^2_1 + \ldots + \sigma^2_m &= \gamma(d_m) \end{aligned}$$

- F-test from ANOVA table; for stage $m+1$: $F = MS_m / MS_{m+1}$

Nested ANOVA : Interpretation

- There is spatial dependence from the closest spacing until the F-ratio is not significant.
- Samples from this distance are independent
- To take advantage of spatial interpolation, must sample closer than this
- Can estimate how much of the variation is accounted for at each spacing

Answers

Q1 : *What is the main purpose of sampling in an area that has never been sampled before? What important property should the sample have?* •

A1 : *A first sample is to establish the* **range of variability** *in feature space of the area, e.g. an empirical distribution of the target variables. It also is used to determine if there is* **spatial dependence** *and if so, its* **structure**; *this can later be used for optimal sampling for mapping.*

The sample should include locations where we might expect differences; also, there must be locations with **different separations** *to estimate a variogram.* *Return to Q1* •

Q2 : *What is the main purpose of additional sampling in a heavily-sampled area? What important property should the sample have?* •

A2 : *A supplementary sample is generally taken to* **increase the precision** *of an estimate or reduce the prediction error of a map.*

Observations should be located where there is maximum uncertainty with the existing sample, or in "hot spots" previously identified by earlier surveys. *Return to Q2* •

Answers

Q3 : *Why should the procedure ever accept a* **worse** *scheme?* •

A3 : *To avoid getting stuck in a* **local minimum**; *sometimes a point should "jump" a long way. Return to Q3* •