Applied geostatistics

Lecture 8 – Spatial sampling

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

March 21, 2014

Topics for this lecture

1. Sampling concepts
2. Design-based spatial sampling
3. *Sample size for design-based sampling*
4. Model-based spatial sampling
5. *Spatial Simulated Annealing*
6. *Sampling for mixed models of spatial dependence*
7. *Nested 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.

Sampling concepts

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.

Spatial sampling refers to a sampling design where the observations are at known **locations**, and the selection of the locations is part of the design.

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 a list of 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 **sampling theory**; some designs require **geostatistical theory** as well.

References

Here are some accessible texts:

- de Gruijter, J.; Brus, D.J.; Bierkens, M.F.P. and Knotters, M., 2006. *Sampling for Natural Resource Monitoring*. Springer.
  - Available as Springer e-book via UT/ITC and other libraries; ISBN 8-3-540-22486-0
- Schreuder, H. T; Ernst, R and Ramirez-Maldonado, H., 2004. *Statistical techniques for sampling and monitoring natural resources*. Gen. Tech. Rep. RMRS-GTR-126. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station; `http://www.fs.fed.us/rm/pubs/rmrs_gtr126.html`
- Cochran, W. G. 1977. *Sampling Techniques* (3rd ed.). New York: John Wiley.
  - Classical design-based sampling theory

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 (usually, a regular grid of) **points**: expected value, variance of estimate, confidence intervals, probability of exceeding a threshold, ...
   - Same, for **block averages**
3. To determine **spatial structure**: direction, range, strength of dependence → causes, processes
4. To **monitor** all of the above over **time** (repeat sampling)

Different statements have **different sampling requirements**

To check your understanding . . .

Q1 : *Suppose we want to map an area that that has never been sampled. What important property should the sample have?* *Jump to A1* •

Q2 : *Why might one want to take additional samples to map an area that has already been sampled? What important property should the sample have?* *Jump to A2* •

Steps in sampling

1. Define the **research questions**
2. Define the **target population**, **target variable** and **target parameter**
3. Define the **quality measure**
4. Specify the **sampling frame**
5. Specify the **sampling design**
6. Determine the **sample size**
7. Determine the **sampling plan**
8. Carry out the **sampling** in the field

Step 1: Research questions

Without knowing **what you want to know** it is impossible to design a sampling scheme to **find out**!

Questions should be precise. Compare these questions related to sampling designs for agricultural research:

- Which is the most widely-grown rice variety in a district?
- What is the total area under rice in the district?
- What proportion of the rice production in a district is from a given variety?
- What was the mean yield of a given rice variety in a given year?
- What is the yield potential of a given rice variety under optimal management?
- What is the relation between yield and soil preparation method? (etc.)

(continued . . .)

And now **spatially-explicit** research questions:

- **Where** are different varieties planted?
- What is the yield **at each location** (e.g., over some fixed grid)?
- Is there any **spatial dependence** between yields, e.g., are there "hot" and "cold" spots?
- If so, what could be the environmental conditions giving rise to these differences?

For all these: How much **certainty** (precision) is needed in the answers?

Step 2 (a): Target population

What exactly is the **population** about which we want to make inferences?

This is the (possibly hypothetical) enumeration of all the **individuals** that make up the population:

- Clear **rules** for **inclusion** or **exclusion** from the sample
- If the population is inherently **continuous** (e.g. "forest cover"), we need a **discretization** rule to divide into individuals.
  - This is the size and shape of the "individual"
  - Equivalent to the concept of geostatistical **support**

The sampling unit

Given the population, we must define it in terms of the **sampling units**.

These are the **individuals** which could be observed or measured.

We must specify:

- How to **identify** (recognize, limit) it in the field;
- How to actually **make the observation** (experimental **protocol**)
  - site preparation for sampling
  - what is to be measured
  - measurement scale and resolution
  - how to measure
  - its **spatial dimensions**, called the **support**

Step 2 (b): Target variable

This is the variable to be **measured** for each sampling unit.

Note that there may be several target variables of interest in the same sampling campaign.

Examples:

- Soil grain size fractions (gravel, coarse sand . . .) in the 0-20 cm and 30-50 cm layers;
- Whether the soil is above a regulatory threshold for some pollutant ("contaminated") or not
- Age of each child in a household and whether s/he attends school regularly

Step 2 (c): Target parameter

This is the **statistical measure** which will summarize the target variable. It is closely related to the research question. What do we really want to know about the **population** from which the **sample** is taken?

Examples:

- Mean proportion of each soil grain size fraction over a study area;
- Mean proportion of each soil grain size fraction of all 1 ha blocks in the study area (**mapping**)
- Minimum, maximum, percentiles ...
- Variance (as a measure of heterogeneity)

These will be estimated by **statistical inference**.

Step 3: Quality measure

Some **numeric measure** of the statistical quality of the inference, i.e., how can we quantify the success of the sample + inference?

Examples (from de Gruijter *et al.* 2006):

- **estimation**: the half-width of a 95% confidence interval of the estimate
- **prediction** (e.g., mapping): the error variance of the prediction
- **hypothesis testing**: the power of the test
- **classification**: the error rate of the classification

Step 4: The sampling frame

This is a technical term for the list or **enumeration** of **all possible sampling units** for the survey.

- This does *not* have to be the population about which the inference will be made, it is often some sub-population which is eligible to be selected in the sample.
- If not the whole population, the researcher must **argue** that it is **representative** of the population; this **meta-statistical** reasoning is used to make inferences about the population from the sample.
- In **spatial** sampling the frame is often a **tesselation** (regular division) into grid cells (squares or hexagons) of the whole area of interest (population) or some representative sub-area.

Example of sampling frame

1. The **population** is all shifting-cultivation fields in the humid tropical rainforest of Cameroon;
   - This is the population about which we want to make some statistically-valid statements.
2. The **sampling frame** includes all shifting-cultivation fields in four "representative" villages;
   - Only these are considered for sampling.
3. The **sample** will be some selection of these fields.

We have to argue (with evidence) that the four villages **represent** the whole area.

We have to ensure that each individual in the sample has a known **probability** of being selected.

To check your understanding . . .

Q3 : *Why would the sampling frame be a sub-population, rather than the whole population?* *Jump to A3* •

Q4 : *Is it statistically-valid to limit the sampling frame to easily-accessible areas, or villages that are known to be cooperative with researchers? Why or why not?* *Jump to A4* •

To check your understanding . . .

Suppose we are designing a field sample to determine soil organic carbon (SOC) stocks over an area covered by a thematic mapper satellite image, of which we will use a vegetation index (e.g., NDVI) as a covariate.

Q5 : *What is a reasonable population, i.e., individuals about which we will make an inference?* *Jump to A5*

•

Q6 : *If the area covered by the image is very large, so that sampling over the whole area is impractical, what would be a reasonable sampling frame?* *Jump to A6* •

Sampling fraction

This is the **proportion** of individuals in the **sampling frame** that are actually selected and sampled. If N is the population size and n is the number of individuals sampled:

$$f = \frac{n}{N}$$

Example 1: In a study area of 100 ha $=$ 1 000 000 m^2; sampling individuals are defined as 10 x 10 m surface areas; so there are $10^6/10^2 = 10^4$ sampling individuals.

If we make 50 observations (e.g. biomass in the 10 x 10 m area) the sampling fraction is $50/10^4 = 0.005 = 0.5\%$.

Example 2: Sampling individuals are households; the sampling frame is the 150 households in a village.

If 20 are selected and interviewed, the sampling fraction is $20/150 = 13.\bar{3}\%$

Further steps

5 Specify the **sampling design** – discussed below ("design-based" and "model-based")

6 Determine the **sample size**

7 Determine the **sampling plan** – the field **logistics** to reach the sampling individuals

8 Carry out the **field sampling**

Spatial sampling

We now consider sampling in space.

First, we must distinguish:

- several views of **spatial structure**; and
- several views of how **randomness** arises in spatially-distributed samples.

Views of spatial structure

There are three conceptual structures of spatial fields:

DMSV **Discrete** model of spatial variation: crisp boundaries between homogeneous units; no spatial structure within units (polygons); widely-separated polygons of the same class ("mapping unit") have the same feature-space distribution of the target value.

- Note there is expected to be variability within polygons, but no spatial structure to this variability.
- Examples: agricultural or forest management parcels; soil mapping units

CMSV **Continuous** model of spatial variation: no boundaries, no units; the target value varies continuously (at some discretization) over space.

MMSV **Mixed** model of spatial variation: there are polygons, grouped in classes, but within these there is continuous variation of the target variable. All polygons of a class have the same internal spatial structure.

Views of randomness

There are two approaches to randomness (which is necessary for valid statistical inference).

Design-based : randomness comes from the known **probability** of including an individual in the sample;

Model-based : randomness comes from an assumed **model** of spatial structure; the **realization** (what is encountered in nature) is viewed as the result of a (spatially-correlated) random process.

.

References

In addition to the de Gruijter *et al.* and Schreuder *et al.* texts listed above, many journal articles deal with spatial sampling, e.g.:

- **Brus, D.J. & 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. & 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.

- **Brus, D. J., & Heuvelink, G. B. M.**, 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma, 138(1-2), 86-95.

(continued . . .)

- **Walvoort, D. J. J., Brus, D. J., & de Gruijter, J. J.**, 2010. *An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means*. Computers & Geosciences, 36(10), 1261-1267.

- **Minasny, B. & McBratney, A.B., 2006**. *A conditioned Latin hypercube method for sampling in the presence of ancillary information*. Computers & Geosciences, 32(9): 1378-1388.

Models of spatial variation – relation to sampling

- **DMSV**: inferences within class estimated by all samples from the class
  - **"Design-based"** sampling, based on feature-space structure (e.g. strata, classes, continuous feature-space predictors)
- **CMSV**: may have 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), especially for
    - determining **spatial structure**; and
    - **mapping** as a continuous field
  - **"Design-based"** can be used here too, especially for **inferences** of (sub)-**population parameters**
- **MMSV**:
  - Stratify by DMSV, model within by CMSV

What is different about designs considering spatial structure?

1. DMSV: No assumptions about spatial structure; the probability-based estimates of (sub)-population paramters are not biased;

2. CMSV: We can place samples for maximum information at 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**.

   - DMSV is often used to determine **strata** and per-stratum sample size; then CMSV is used for model-based sample placement in space.

Design-based spatial sampling

In the Discrete Model of Spatial Variability (**DMSV**) we assume all locations in a stratum have the same probability distribution of the target variable; therefore there is no need to model spatial dependence. Sampling is **design-based**.

Even if we know there is spatial dependence, a design-based sample is still valid and preferred for some inferences.

To check your understanding . . .

Q7 :

(1) In the DMSV, is it necessary that all locations in a stratum have identical values of the target variable?

(2) What must be "the same" about all locations within a stratum for proper statistical inference in the DSMV?

Jump to A7 •

Sampling designs with the DMSV

This is classical statistics applied to sampling units that are located in geographic space, but where this is not considered a predictive factor.

- Consider the coordinates only as identifiers of candidates in the sampling frame;
- May consider predictive factors in **feature space**; leads to **stratified** sampling designs;
- Can use "classical" inference, e.g. multiple regression in feature space assuming no correlation between residuals.

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).

Design-based 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
   - randomness comes from random grid origin
3. **Partitioned**
   - Two stages

   (a) coarse grid of blocks covering the area, select blocks at random;

   (b) within blocks, either completely random or grid

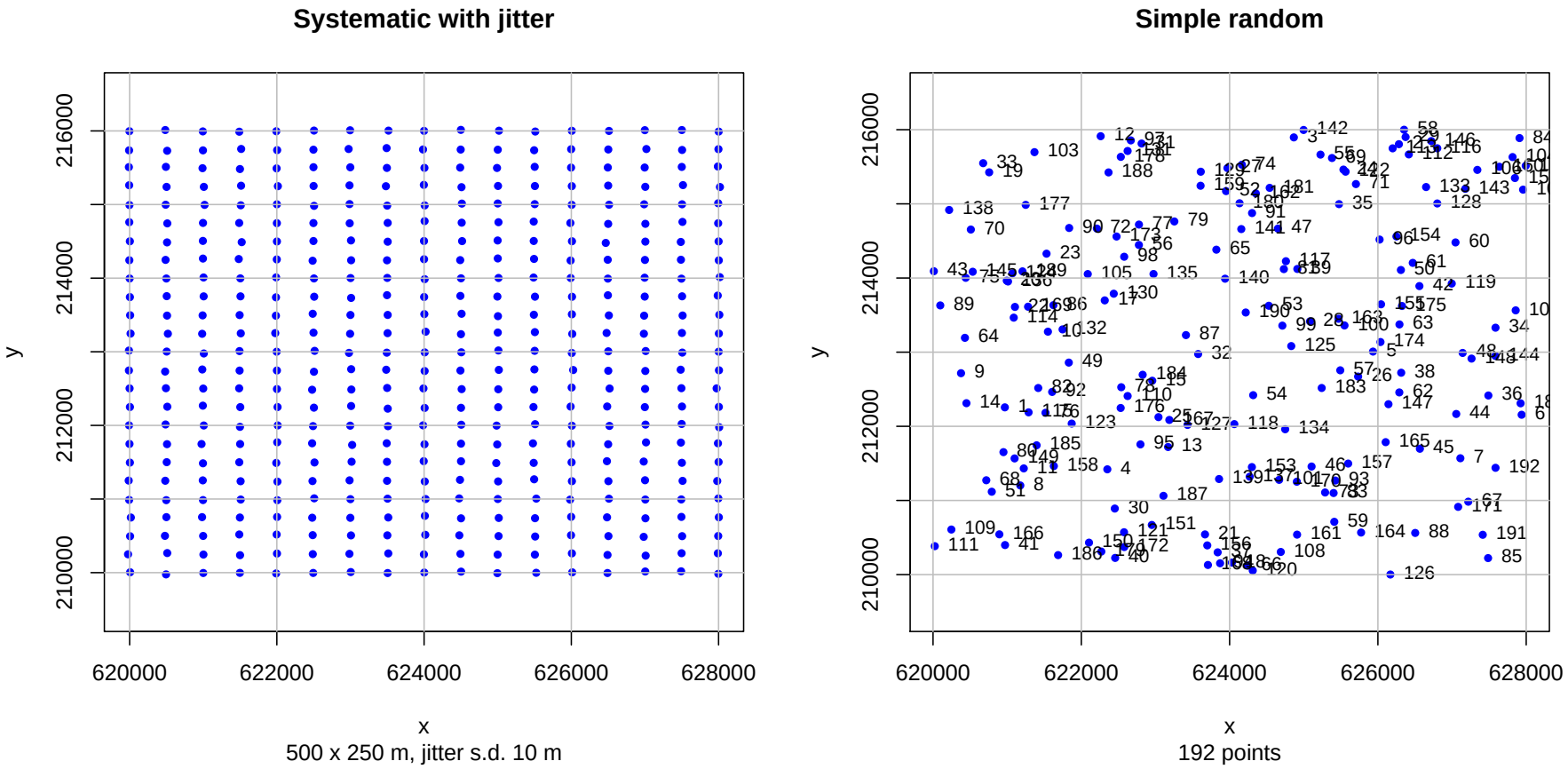
Logistics

Completely random most difficult to move between locations, complex navigation and impossible-to-optimize ("travelling salesman" problem) route planning

Grid easy navigation, less total travel, but longer distances between locations

Partitioned the only long distances are between blocks, within blocks less travel time

Two unstratified sampling designs



Stratified sampling

Often we have prior evidence that the whole space is not a homogeneous unit; we can then make inferences about sub-spaces ("strata") either by:

1. Design-based unstratified sampling (as above), but recording the discriminating variables as **factors** along with the target variables (e.g., land-use class along with biomass);

2. **Stratified** sampling: dividing the area into strata as part of the design.

(continued . . .)

Stratification

- There are **sub-areas** where there is evidence that the target variables should be different
  - In a soil survey the underlying lithology as shown on a **geologic map** is expected to influence the soil texture; and the land-use history as shown on a **land-use map** is expected to influence the soil organic matter;
  - Map units of these maps can be used as strata
- There are continuous **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)
  - These continuous attributes can be classified ("sliced") to produce strata.

How to divide the sample among strata?

1. **Proportional** to the area covered by the stratum: **proportional** stratified sampling;

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 the target variable.

3. **According to interest** in the stratum.

   - More will be known (e.g., lower estimation variance) in the intensively-sampled areas.
   - E.g., soils near industrial sites are expected to be more polluted than rural soils; if the target variable is heavy-metal concentration, we are more interested in the areas expected to have high values.
   - E.g., soils in agricultural areas will be more intensively used than in conservation areas, so the properties of ag. soils may be of more interest.

Topic: Sample size for design-based sampling

Sampling is expensive, but so is incorrect or imprecise information. These two must be balanced by determining the **sample size** that will satisfy information needs while minimizing costs.

We first illustrate the concept of sampling error, then develop theory to determine sample size, then see how to compute it:

1. Sampling error
2. Theory
3. Computation

Sampling error

Estimates from samples are almost never equal to true values, and estimates from different samples differ among themselves.

To quantify this we define the concept of **sampling error**:

- The amount by which an **estimate** of some population parameter computed from a **sample** deviates from the **true value** of that parameter for the **population**.

Example: Estimated total rice production in a district, extrapolated from a sample of fields, vs. the actual total production.

Of course we usually don't know this (since we don't know the true value).

Sampling error

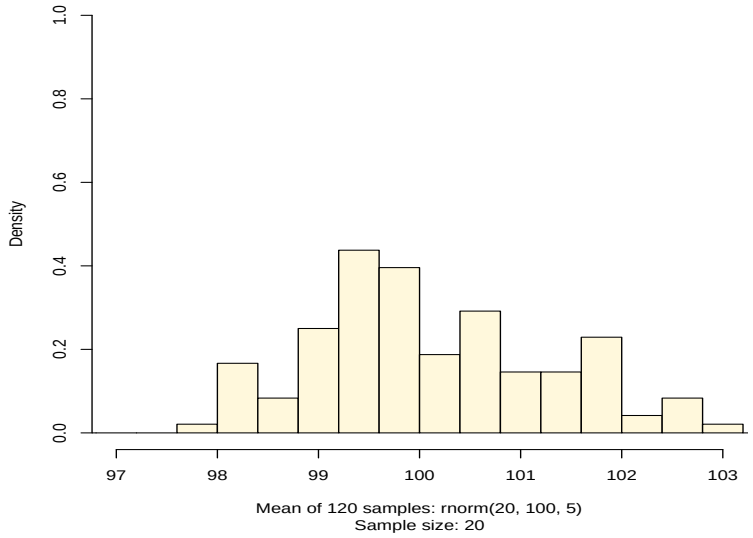
We can appreciate sampling error by **simulation** from known populations.

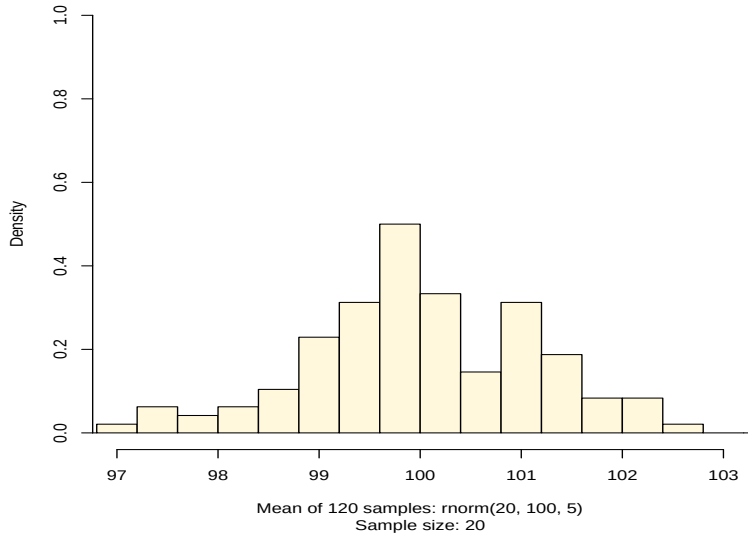
Example: Draw 10 different random samples from a normal distribution with true mean 100 and standard deviation 5; size of each sample is 20 observations; compute the sampling errors:

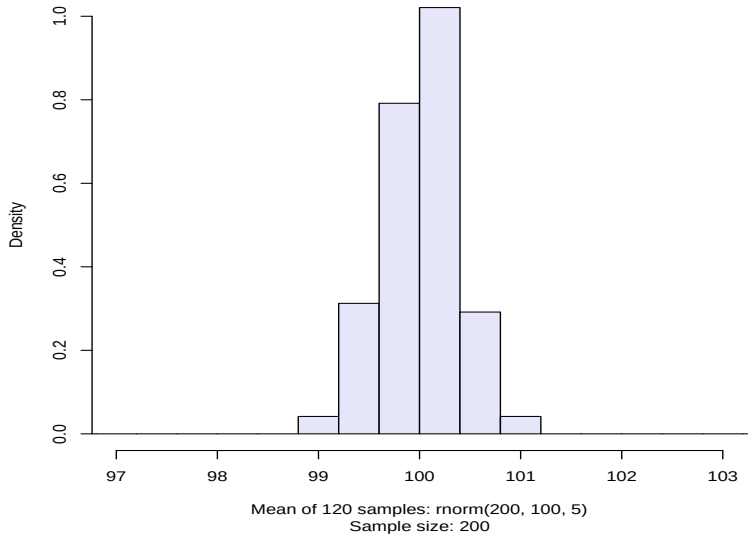
```
R> # set up a vector for the results
R> samp <- rep(0, 10)
R> # compute the means of 10 sets of 20 normal variates, true mean=100
R> for (i in 1:10) samp[i] <- mean(rnorm(20, 100, 5))
R> # compute sampling errors
R> 100 - samp
 [1] -1.480606  0.256055  0.165392 -0.096576 -0.730931 -0.109797
 [7]  1.118741 -0.246498 0.674641  0.887922
R> # mean sampling error
R> mean(100 - samp)
[1] 0.043834
```

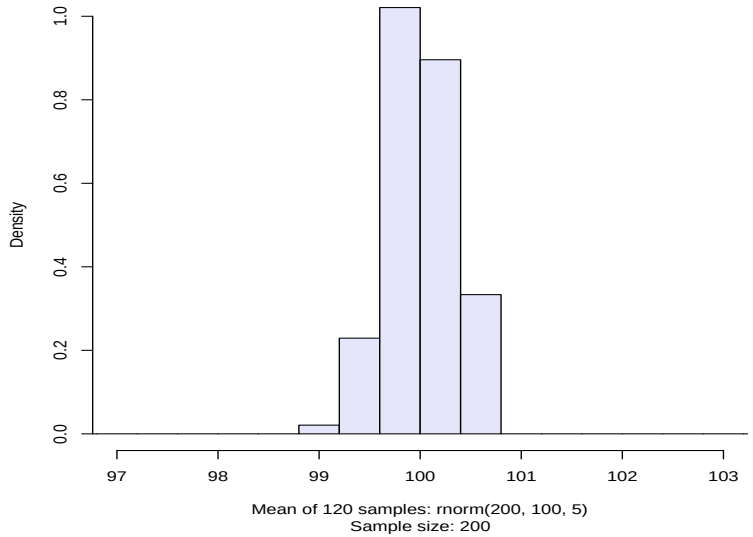
Notice that the **mean** sampling error is almost zero. This is the result of the **central limit theorem**, derived from the **law of large numbers**.

Illustrating the central limit theorem









Type I and Type II error

To understand how we determine sample size, we need to recall some basics of **hypothesis testing**.

There are two types of inferential errors we might make:

Type I : **rejecting** the null hypothesis when it is in fact **true**; a **false positive**

Type II : **not rejecting** the null hypothesis when it is in fact **false**; a **false negative**

	Null hypothesis H_0 is really ...	
Action taken	True	False
Reject	Type I error committed	success
Don't reject	success	Type II error committed

Significance levels

There are two risk levels associated with the two types of error:

α is the risk of **Type I** error

We set α to guard against false inference; thus we are inherently **conservative**.

β is the risk of **Type II** error

$1 - \beta$ is known as the **power** of the test (see below).

We get β from the form of the test and true effect (see below).

Example

Null hypothesis: A new crop variety will not yield at least 100 kg ha^{-1} more than the current variety; that is, there is no real reason to recommend the new variety.

Note: this is an **informative** null hypothesis; not just "no difference". It is set by the researcher. In this case, unless we can prove this much difference we won't bother to develop the new variety. This is a management decision, not statistical.

Type I error: the new crop variety in fact **does not** have an average yield (if grown "everywhere") at least 100 kg ha^{-1} more than the current variety, but from our (limited) sample we say that it **does**. A "false positive". So, we develop the variety and recommend it, but the farmer gets no significant benefit.

Type II error: the new crop variety in fact **does** have an average yield (if grown "everywhere") at least 100 kg ha^{-1} more than the current variety, but from our (limited) sample we say that it **does not**. A "false negative". So, we abandon the variety, even though the farmer would have benefitted.

Approaches to computing sample size

There are two main approaches:

1. Power analysis

2. Sampling to narrow a confidence interval

For most applications in spatial sampling, we are interested in the **confidence interval** of an estimate, and so will use the second approach.

Power analysis is commonly used to design experiments with a known probability of revealing differences between treatments.

Note: See the website
`http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3`
(`G*Power 3` computer program) for this approach.

Sampling to narrow a confidence interval

One approach to sample size calculation is to consider the desired **width of the confidence interval** for some parameter of interest.

We will use the example of a **confidence interval for a mean value**.

Confidence interval for the mean

Recall that the **confidence interval** for a mean μ is computed as:

$$(\bar{x} - t_{\alpha/2,n-1} \cdot s_{\bar{x}}) \leq \mu \leq (\bar{x} + t_{\alpha/2,n-1} \cdot s_{\bar{x}})$$

where:

- $\bar{x}$ is the sample mean;
- $t_{\alpha/2,n-1}$ is Student's t with $n-1$ degrees of freedom at confidence level $\alpha/2$;
- $s_{\bar{x}}$ is the standard error of the sample mean:

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

$$s_x = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$$

Notes on these formulas

- The confidence level α, say 0.05, is halved, say to 0.025 for each side of the interval, because this is a two-sided interval.

- The **t-distribution** must be used because we are estimating both the mean and variance from the same sample; for reasonably-large sample sizes the **normal distribution** itself (here called the z distribution) can be used.

What affects the confidence interval?

1. The t value:
   (a) n: **sample size**: $t \rightarrow z$ as $n \rightarrow \omega$
   (b) α: **risk level** set by the experimenter that the computed interval does **not** contain the true mean; a higher risk leads to a narrow interval

2. The standard error $s_{\bar{x}}$:
   (a) n: **sample size** (again): precision increases as $\sqrt{n}$
   (b) the **sample standard deviation**: this is essentially the **inherent variability** of the sample

What can we control?

1. The **risk** of rejecting a true null hypothesis (α); depends on the **cost of a false positive**;
   - If there is little cost associated with making a Type I error, α can be high (lenient); this will narrow the confidence interval.

2. **Sample standard deviation**
   - We have some control by good experimental or observational procedures
   - But we can not control the **inherent variability** in the population, even with perfect technique;

3. The **half-width** w of the confidence interval, i.e. the required **precision**. We set this according to how precise the computed estimate must be; this depends on the application.

Inverting the confidence interval

With the above parameters set, we can compute the required **sample size**:

1. **Set** the required risk α that the computed mean value (or mean difference) is outside the interval;

2. **Set** the desired (half-)**width** of the confidence interval w;

3. **Estimate** the **sample standard deviation** s_x

Then we **solve for** n:

$$
\begin{aligned}
w &= t_{\alpha/2,n-1} \cdot s_x / \sqrt{n} \\
\sqrt{n} &= t_{\alpha/2,n-1} \cdot s_x / w \\
n &= (t_{\alpha/2,n-1} \cdot s_x / w)^2
\end{aligned}
$$

No closed-form solution for n

There is a problem with this "solution" for n:

The right-hand side also contains n (which we want to compute), because the t-value depends on the degrees of freedom.

So we must somehow **approximate** t, solve, and then **iterate**. In practice either of the following two methods can be used:

1. Replacing t with z: for **larger** expected sample sizes

2. Use a conservative estimate of t: for **smaller** expected sample sizes

Solution 1: Replacing t with z

Replace $t_{\alpha/2,n-1}$ with $z_{\alpha/2}$, i.e. the normal deviate ($= t$ with infinite d.f.).

This leads to **under-estimation** of n by a factor f of (example for $\alpha = 0.05$):

n	10	20	50	100	200	500
f	0.268	0.126	0.049	0.024	0.012	0.005

(Note: R code for this: `qt(0.975, n) - qnorm(0.975)` etc.)

So the sample size estimate will be somewhat **too low**.

Then **iterate** with this first estimate of n, using the t value this time.

Solution 2: Use a conservative estimate of t

Use a small but realistic value of n to compute $t_{\alpha/2,n-1}$; as long as the computed n is larger, this is a conservative estimate.

If a more exact n is needed, the new estimate can be used to re-compute $t_{\alpha/2,n-1}$, **iteratively**.

How to set these values?

1. We **set** the desired **risk** based on how often we are willing to be wrong (i.e. the actual value is outside the computed limits).

2. We **set** the desired **width** based on the **precision** we require.

3. We **estimate** the **sample standard deviation** from a previous study on this sampling frame, or from similar studies. This of course may not be the actual sample standard deviation we get from the new sample.

Numeric example: Problem

1. **Problem**: Determine sample size for a natural resources survey to detect difference in soil carbon stocks between two land use systems (e.g. conventional vs. organic agriculture).

2. The **population** is all fields in either land use system (note – this must be carefully described).

3. We use a **paired** design: sets of adjacent fields, as similar as possible in soils and other management, and measure the soil carbon in each field by some field and lab. protocol.

4. We then compute the **paired difference** and, from these, the **mean difference**.

5. We compute the **confidence interval** of the **mean difference** based on our sample.

Use of the confidence interval

- If this interval includes 0 we can not reject the null hypothesis H_0 of no difference in soil carbon stocks between systems.

- But we get more information here: the **interval** in which the **true difference** is expected to lie: i.e. an estimate of the **magnitude of the effect**. This can be directly used for decision-making.

Numeric example: Setup

1. We set the **risk of rejecting a true null hypothesis** α to 0.1 because we are willing to accept a 10% risk of falsely rejecting the null hypothesis (i.e. falsely deciding that one of the alternatives is better than the other). So for each half-width we use half of this, i.e. 0.05.

2. We set the **half-width** to 0.5 kg m^{-2} surface area, because a smaller carbon difference is not considered important for soil behaviour.

3. From a previous survey we estimate the **population standard deviation** to be 2 kg m^{-2}; note that this will be higher with on-farm trials than in controlled experiments.

Numeric example: Solution

We begin with an estimated sample size of 20; we know this is within our budget.

$t_{.05,19} = 1.7291$ (R code: `qt(.95, 19)`)

$$
\begin{aligned}
n &= (t_{\alpha/2,n-1} \cdot s_x / w)^2 \\
n &= (1.7291 \cdot (2/0.5))^2 \\
n &= 47.8 \approx 48
\end{aligned}
$$

This suggests that a sample size of 48 should detect a real difference of 0.5 kg m^{-2} in either direction, with a risk of 10% of incorrectly calling a chance difference real.

Note this is 48 pairs, since it is a paired test.

Numeric solution: iteration

Now that we know $\approx n$ we can recompute t and refine the estimate; with this higher n the t value will be a bit lower and so will the required sample size.

$t_{.05,47} = 1.6779$ (R code: `qt(.95, 47)`)

$$
\begin{aligned}
n &= (t_{\alpha/2,n-1} \cdot s_x/w)^2 \\
n &= (1.6779 \cdot (2/0.5))^2 \\
n &= 45.047 \approx 45
\end{aligned}
$$

The difference is small, only 3 fewer samples.

Numeric solution: normal approximation to t

$z_{.05} = 1.6449$ (R code: `qnorm(.95)`)

$$
\begin{aligned}
n &= (z_{\alpha/2} \cdot s_x / w)^2 \\
n &= (1.6449 \cdot (2/0.5))^2 \\
n &= 43.289 \approx 43
\end{aligned}
$$

This is only two fewer than when using the correct t, so it is also a good approximation. We could now iterate with $t_{0.05,42}$ as above, and arrive at the final (correct) sample size, $n = 45$.

Effect of parameters

The following all **increase** the required sample size:

1. α: $\alpha = 0.10 \rightarrow 45$; $\alpha = 0.05 \rightarrow 65$; $0.01 \rightarrow 115$

   Decreasing risk (i.e. a smaller α)

2. w: $w = 1 \rightarrow 11$, $w = 0.5 \rightarrow 45$, $w = 0.25 \rightarrow 180$

   Detecting a **small (real) difference**

3. s: $s = 1 \rightarrow 11$, $s = 2 \rightarrow 45$, $s = 4 \rightarrow 180$

   A **more variable population**

Model-based spatial sampling – overview

Continuous Model of Spatial Variability (**CMSV**): the only structure is **spatial**.

I.e., we assume that the target attribute is the result of a **random field** which structure must be **modelled**; thus the sampling must be adequate to build this model.

Since we hypothesize a spatially-correlated CMSV, sampling has two main aims:

1. Determine this **spatial structure** (e.g., fit a variogram model);

2. **Map** the spatially-correlated field by interpolation with the modelled spatial structure (e.g., by Ordinary Kriging)

Commentary

If the analysis of spatial structure is not successful, it is impossible to map by the CMSV; the only honest approaches are Theissen polygons if there is evidence of continuity or the overall mean otherwise.

Sampling designs to determine spatial structure

Here the aim is to determine spatial structure:

- as an end in itself (to understand landscape processes);
- to build a model to be used in kriging interpolation.

The sampling scheme is *not* necessarily optimal for mapping an area; in fact the requirements are quite different:

- Determine spatial structure: sample at various resolutions;
- Mapping: spread observations over the entire area.

Thus there may be large "holes" in the coverage – but these are assumed to have the same spatial structure as the areas where observations are located.

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.

Number of observations to model the variogram

Stochastic simulation from a random fields with known variograms suggests:

1. $<$ 50 points: not at all reliable
2. 100 to 150 points: more or less acceptable
3. $>$ 250 points: almost certaintly reliable

More points are needed to estimate an **anisotropic** variogram

Reference:

- **Webster, R., & Oliver, M. A.** (1992). Sample adequately to estimate variograms of soil properties. *Journal of Soil Science*, 43(1), 177-192.

Sample spacing – widest

We must have some *a priori* idea of the range of spatial dependence, i.e., the maximum separation at which there is expected to be any dependence – is it 100 km, 10 km, 1 km, 100 m, 10 m, 1 m, . . . ?

Spacings wider than this will not contribute to discovering the model of spatial dependence.

Spacings that do not reach this will leave part of the spatial structure un-discovered.

This *a priori* range may be inferred from:

- previous studies of the same target variable in similar areas;
- ranges of covariates.

To check your understanding . . .

Q8 : *Suppose the aim is to map target variable "biomass" of a forested area.*

Further suppose we have available a thematic mapper satellite image, from which we can easily computevegetation indices (e.g., NDVI) which are known to be related to vegetation vigour.

How could we use the image to estimate the a priori range of a variogram for the target variable? *Jump to A8* •

Sample spacing – narrowest

This depends on the narrowest separation where information is needed.

For eventual mapping, this would be the interpolation block size.

Another way to think about it: what is the closest spacing within which you are willing to be ignorant of local variability?

Note: narrower separation → lower nugget, therefore lower minimum kriging variance at that block size.

Placement of observations

One approach is the nested scheme of Webster; see separate topic below.

Another approach is the **geometric-series transect**:

- Transect: place lines across landscape, establish sampling points along these
- If anisotropy is suspected, transects along and across main variation (orthogonal)
- Within each transect, a **geometric series** of spacings

Geometric series

This most **efficiently** captures a **wide range of variability**.

Boundary conditions: widest spacing: s_1; narrowest spacing: s_2

$s_3 = \sqrt{s_1 \cdot s_2}$; just one of these

$s_4 = \sqrt{s_1 \cdot s_3}$; $s_5 = \sqrt{s_2 \cdot s_3}$: two of these, divides each interval

Can continue to get finer resolution (more variogram bins, so better variogram modelling).

Tradeoff: more stations in each transect → fewer transects (with same total number of observations)

Geometric series: example

- Boundary conditions: $s_1 = 600$ m (widest), $s_2 = 6$ m (closest)
- First intermediate spacings $s_3 = \sqrt{6 \text{ m} \cdot 600 \text{ m}} = 60$ m
- Series now {600m, 60m, 6m}
- Second set of intermediate spacings, as the geometric mean between each adjacent spacing:
  - $s_4 = \sqrt{600 \text{ m} \cdot 60 \text{ m}} \approx 190$ m
  - $s_5 = \sqrt{60 \text{ m} \cdot 6 \text{ m}} \approx 19$ m
- Final series {600 m, 190 m, 60 m, 19 m, 6 m}

Placement on the transect

In the previous example, the transect is 600 m long, or else divided into 600 m segments; observations are made at each end.

Then from one end, an observation is placed 190 m from that end.

Then from that end and the new station, observations are placed 60 m from these, in the same direction.

This is repeated for the 19 m and 6 m spacings.

The final layout is:

0, 6, 19, (19+6)=25, 60, (60+6) = 66, (60+19) = 79, (60+19+6) = 85, 190, (190+6) = 196, (190+60) = 250, (190+60+6) = 256, (190+60+19) = 269, (190+60+19+6) = 275, 600.

When the spacing gets quite close (here, 6 m) it may be possible to omit some (e.g., half) of these spacings.

Commentary

This could be extended with a third set of intermediate spacings, as the geometric mean between each adjacent spacing:

- $s_6 = \sqrt{600\ \mathrm{m} \cdot 190\mathrm{m}} \approx 338\ \mathrm{m}$
- $s_7 = \sqrt{190\ \mathrm{m} \cdot 60\mathrm{m}} \approx 107\ \mathrm{m}$
- $s_8 = \sqrt{60\ \mathrm{m} \cdot 19\mathrm{m}} \approx 34\ \mathrm{m}$
- $s_9 = \sqrt{19\ \mathrm{m} \cdot 6\mathrm{m}} \approx 11\ \mathrm{m}$

Final series $\{$600 m, 338 m, 190 m, 107 m, 60 m, 34 m, 19 m, 11 m, 6 m$\}$

Sampling designs for mapping with the CMSV

A main use of the CMSV concept is to map an area as a "continuous" surface – in practice, it is some **tesselation**.

Desiderata:

1. Maximize information
   - Cover the largest possible area
   - Minimize some **optimization criterion** in the resulting map
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.

Note that if there are areas that need different precisions, the numerical value of the quality measure can differ between them.

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 may be **perturbed** so samples do not line up exactly; avoids unexpected periodic effects

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

Commentary

Some practicioners add a small random **jitter** to each location, to avoid undected linear features This alters the inclusion probabilities in a design-based scheme, but in a model-based scheme randomness does not arise from observation locations, so there is no need for known inclusion probabilities.

Optimal point configuration in the presence of anisotropy

Optimal designs are easily adjusted for **affine** (also called **geometric**) anisotropy.

This is where the range of spatial dependence differs for two othogonal axes, but the variogram sill, and model form are the same along both axes.

Adjustment : stretch it in the direction of maximum dependence, based on the **anisotropy ratio**, i.e., the ratio of the two ranges.

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 with a known variogram model

- 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)
- Note that the **values** of the target variable are nowhere in this formula!
- 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.

To check your understanding . . .

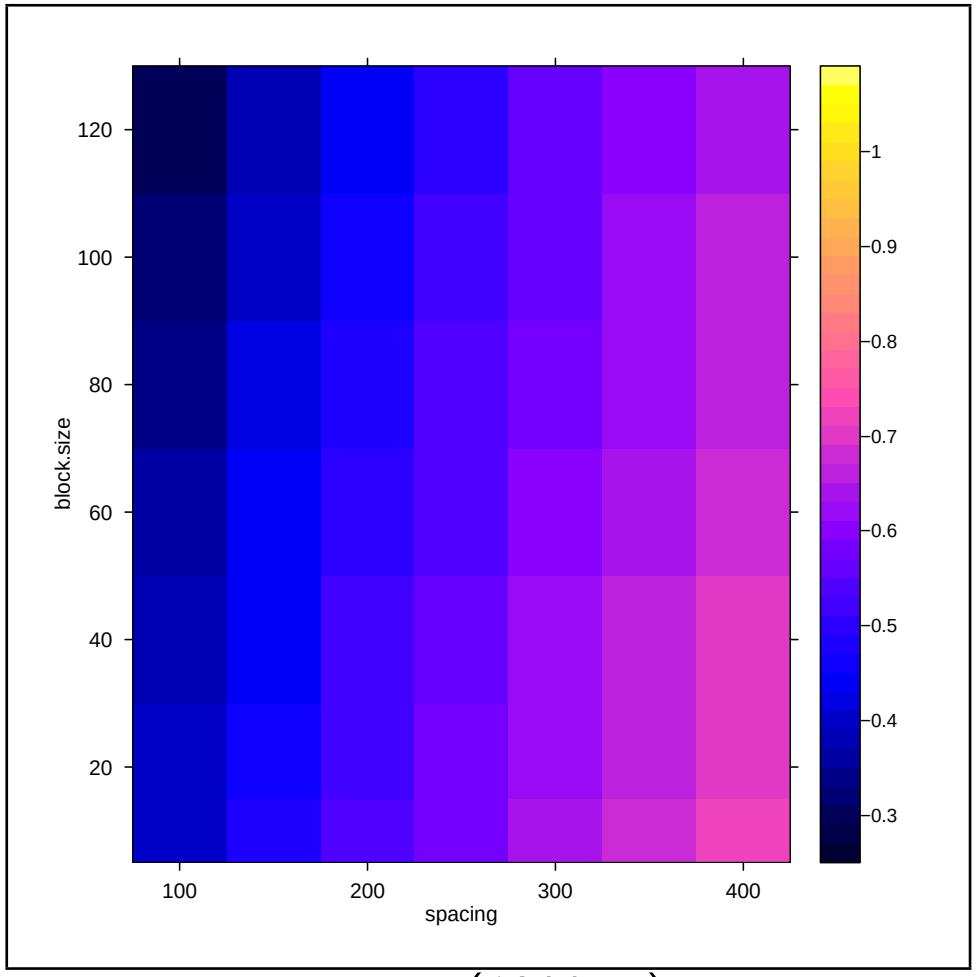
Q9 : *What information is lost as the prediction block size increases?* *Jump to A9* •

Q10 : *Suppose the target variable is the concentration of some hazardous soil pollutant, which will be removed if its concentration is predicted to exceed some threshold. What is the public-health danger of using too large a block size?* *Jump to A10* •

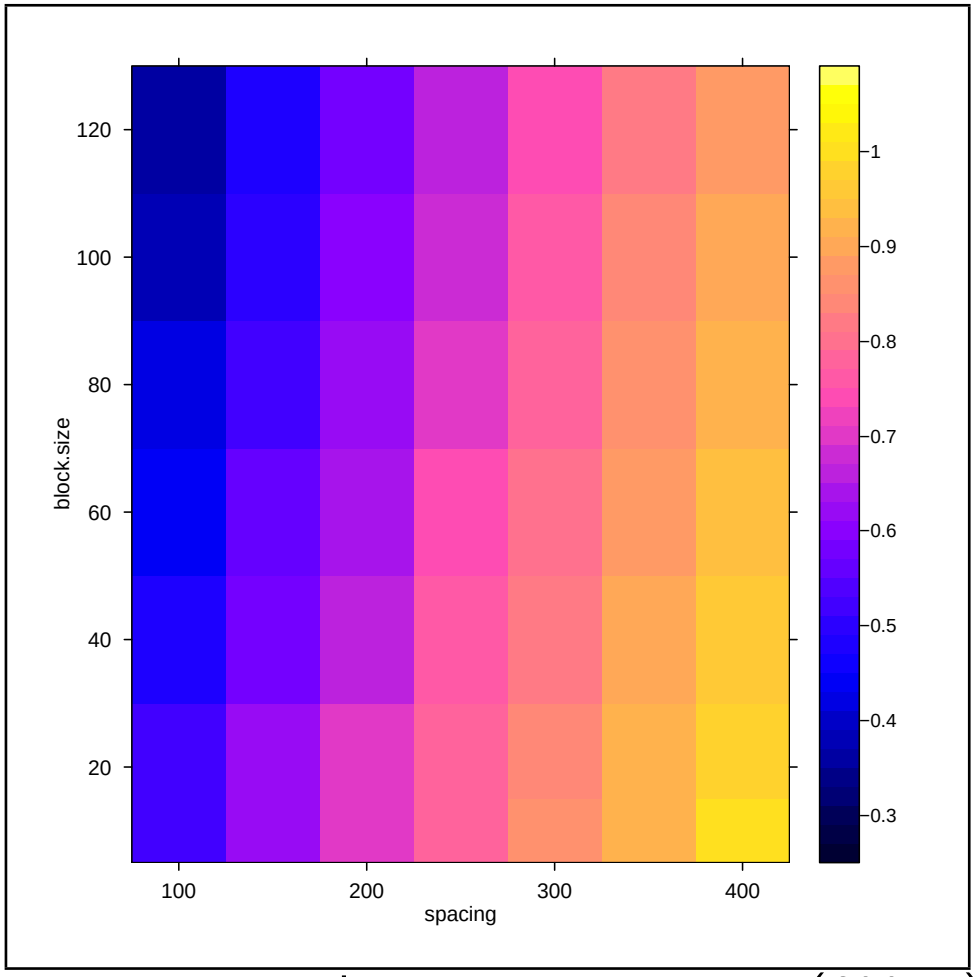
Error as a function of increasing grid resolution

- 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)

To check your understanding . . .

Q11 : *Considering the long-range variogram, what spacing between observations would be needed to obtain the same or lower kriging variance for a 20 m block as using a 400 m spacing and a 120 m block?* *Jump to A11* •

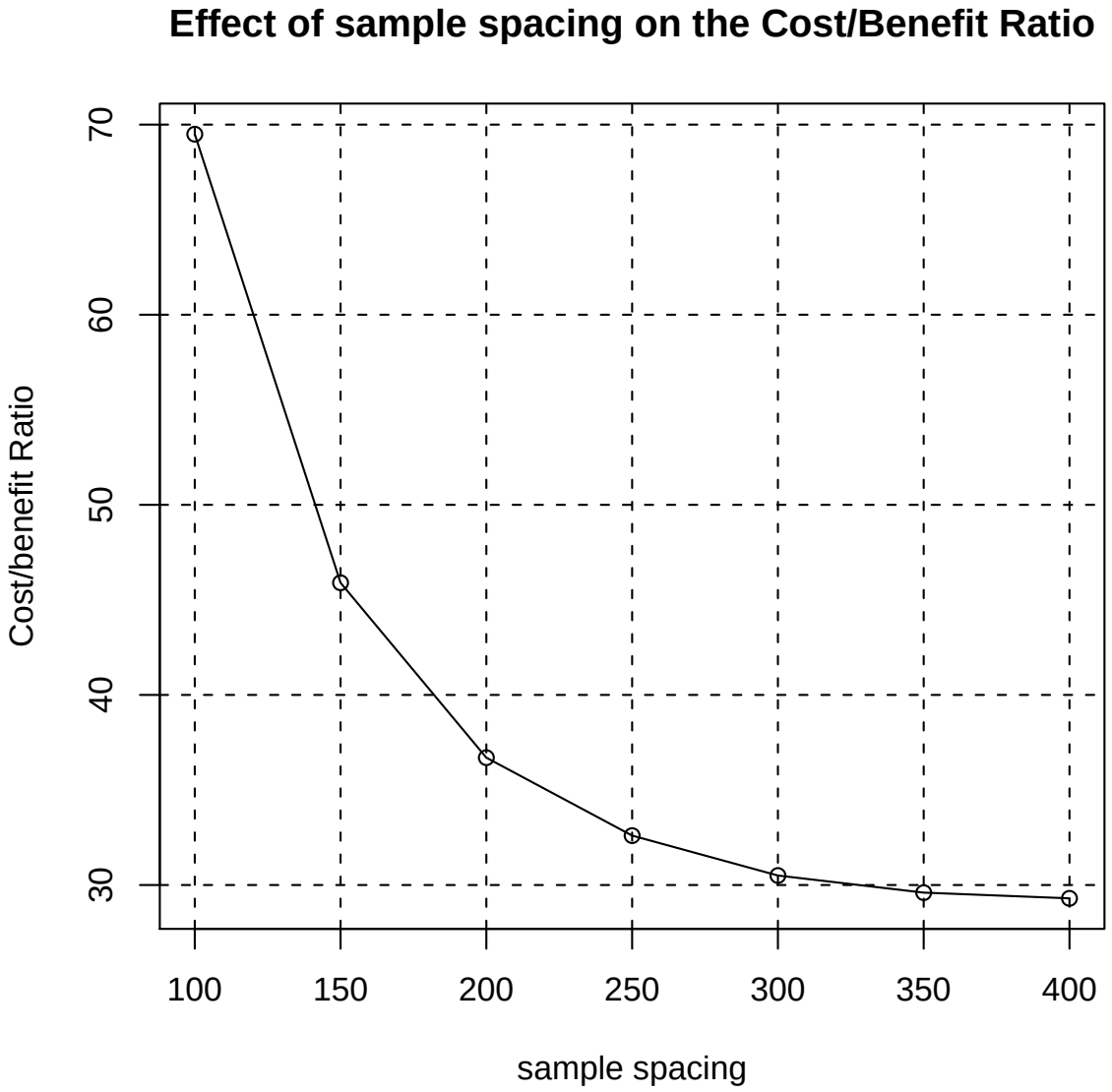
Q12 : *If now the variogram is short-range (i.e., the spatial dependence is only over shorter separations), what is the widest spacing that could be used to obtain the same or lower kriging variance as using a 400 m spacing and a 120 m block for the long-range variogram?* *Jump to A12* •

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: 25 km x 25 km 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**):

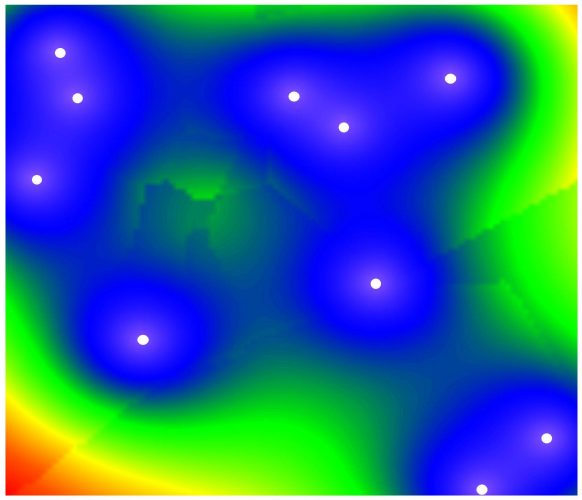
   (a) Select a point to move; **move** it a **random distance and direction**
   (b) If outside study area, try again
   (c) Compute **new fitness**
   (d) If **better**, accept new plan; if **worse** also accept with a certain **probability**

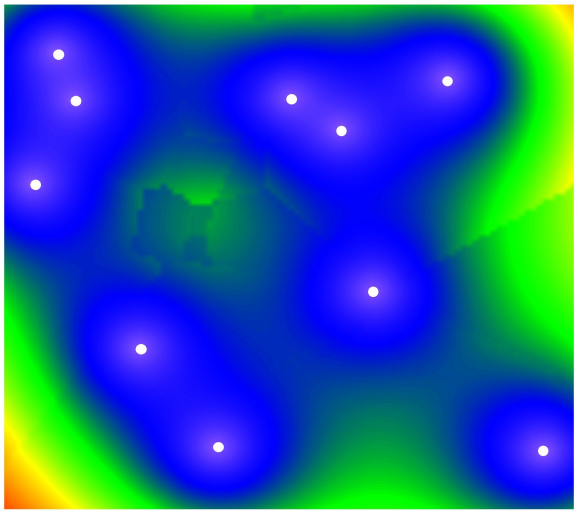
4. **Stop** according to some **stopping criterion**

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:

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

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:

$$p = e^{-\Delta f / T_k} \tag{3}$$

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

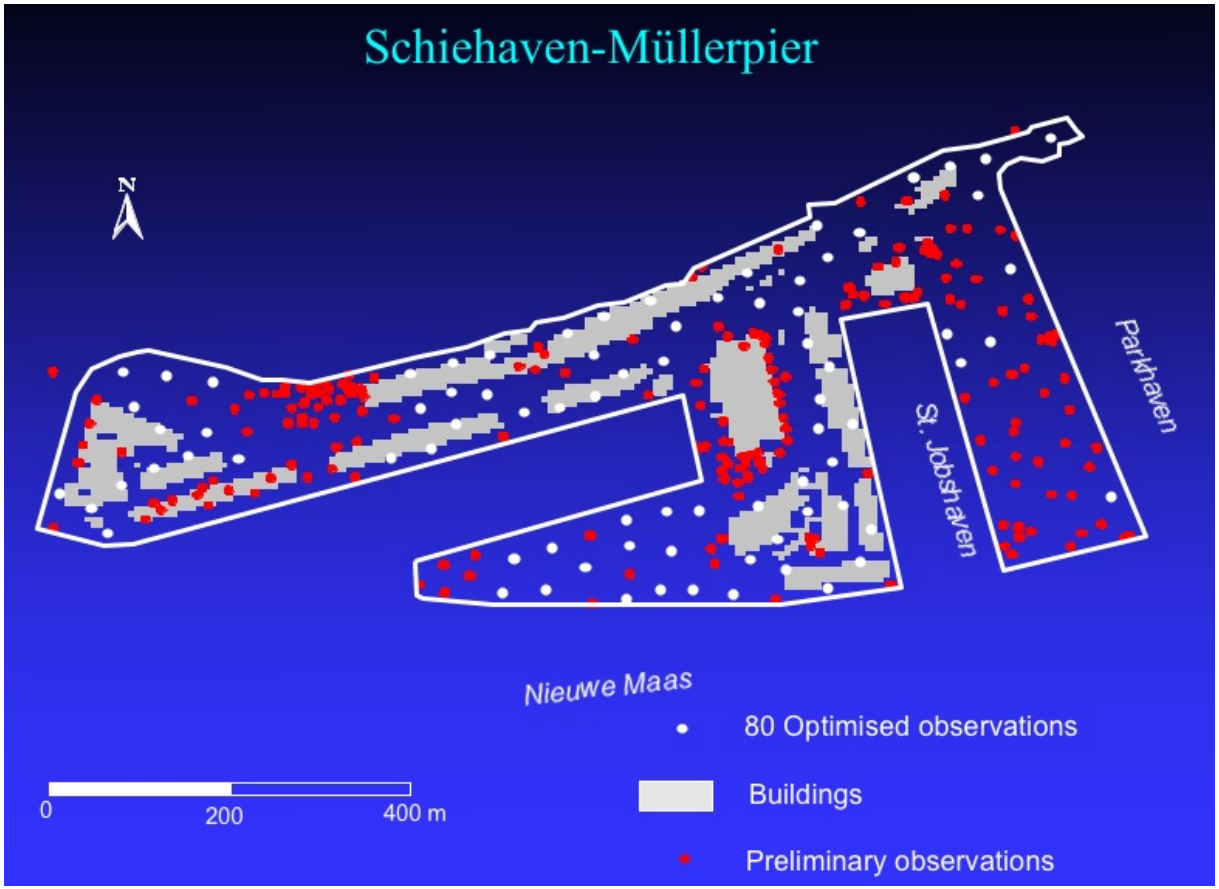
Note that this will be positive for a poorer solution, so its complement is used for the exponent.

To check your understanding . . .

Q13 : *Why should the procedure ever accept a* **worse** *scheme?* *Jump to A13* •

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: Use design-based inference.

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

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.

Purpose: establish the structure of spatial dependence (e.g. variogram) with a minimal number of samples.

Not intended to map an area, although the fitted variogram that results can be used to design an optimal sampling scheme to map (e.g. OSSFIM).

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

Nested sampling

- Various spacings between observations, designed in stages
- 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
- Fill in the series with a **geometric series**, each series "nested" within previous ones.

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)

Example of nested sampling

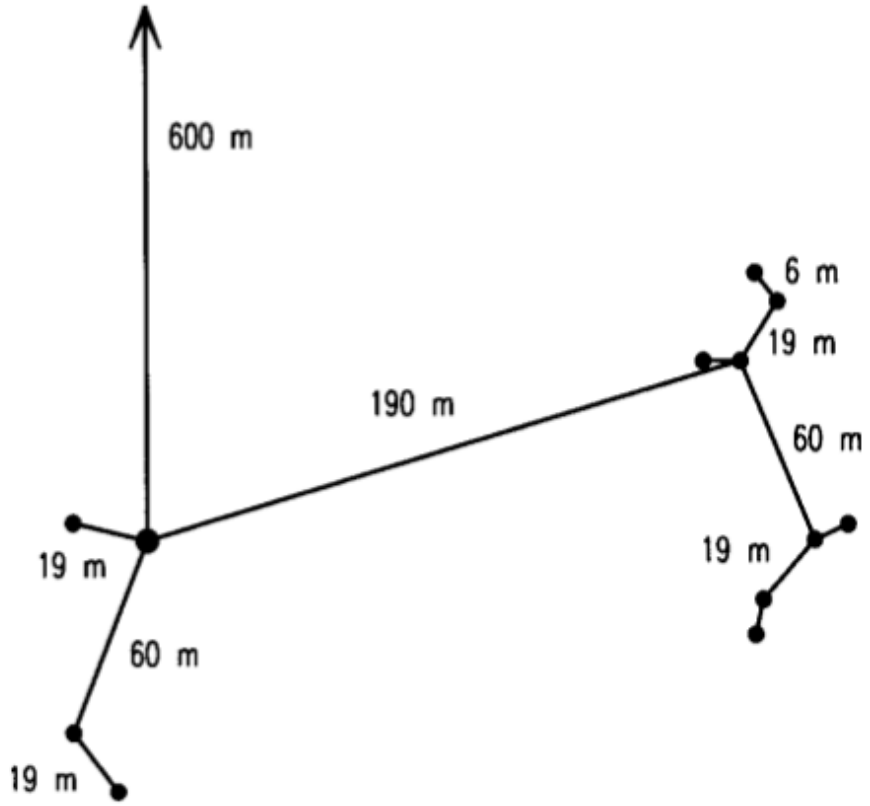


Figure 6.14 Sampling plan for one main centre of the Wyre Forest survey (not strictly to scale).

Source: **Webster, R., and M.A. Oliver** (2008). Geostatistics for environmental scientists. 2 ed. John Wiley & Sons Ltd.

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 : *Suppose we want to map an area that that has never been sampled. What important property should the sample have?* •

A1 : *A first sample should 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 : *Why might one want to take additional samples to map an area that has already been sampled? 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 would the sampling frame be a sub-population, rather than the whole population?* •

A3 : *The main reason is usually* **logistics***: it is impractical to travel to all possible sampling individuals in the population. In the Cameroon example access in the rainforest zone is slow and uncertain, so it makes sense to* **cluster** *by selected villages.* *Return to Q3* •

Q4 : *Is it statistically-valid to limit the sampling frame to easily-accessible areas, or villages that are known to be cooperative with researchers?* •

A4 : *This is only valid if we can argue from other evidence that the limited frame represents the population. This seems quite doubtful in the two cases mentioned: accessible land by definition has different management potential (and therefore almost always different use) and often different land characteristics (e.g., gentle slopes, deep soils . . .); cooperative villages obviously have a different dynamic towards researchers, so answers would be expected to be different from un-cooperative villages.* *Return to Q4* •

Answers

Suppose we are designing a field sample to determine soil organic carbon (SOC) stocks over an area covered by a thematic mapper satellite image, of which we will use a vegetation index (e.g., NDVI) as a covariate.

Q5 : *What is a reasonable population, i.e., individuals about which we will make an inference?* •

A5 : *All soil areas covered by one pixel of the covariate image.* *Return to Q5* •

Q6 : *If the area covered by the image is very large, so that sampling over the whole area is impractical, what would be a reasonable sampling frame?* •

A6 : *All soil areas covered by a "representative" sub-image. This could be selected by randomly locating one corner. In practice it may be selected for logistic reasons (accessibility), although the analyst must argue the purposively-chosen sub-image is representative of the whole.* *Return to Q6* •

Answers

Q7 :

(1) In the DMSV, is it necessary that all locations in a stratum have identical values of the target variable?

(2) What must be "the same" about all locations within a stratum for proper statistical inference in the DSMV? •

A7 :

(1) No, there can well be different values at different locations. For example, topsoil sand proportion could vary within a "homogeneous" soil map unit.

(2) But, the **expected value** *and associated* **variance** *(in general, the* **probability distribution***) of the target variable must be same everwhere – until we sample at a specific location, we have the same expectation and same uncertainty.* *Return to Q7* •

Answers

Q8 : *Suppose the aim is to map target variable "biomass" of a forested area.*

Further suppose we have available a thematic mapper satellite image, from which we can easily computevegetation indices (e.g., NDVI) which are known to be related to vegetation vigour.

How could we use the image to estimate the a priori range of a variogram for the target variable? •

A8 : *Compute and model a variogram of the NDVI over the image; the range of the modelled variogram is a reasonable estimate of the range of biomass.* *Return to Q8* •

Answers

Q9 : *What information is lost as the prediction block size increases?* •

A9 : *All information about the distribution of the target property* **within** *the block is lost.* *Return to Q9* •

Q10 : *Suppose the target variable is the concentration of some hazardous soil pollutant, which will be removed if its concentration is predicted to exceed some threshold. What is the public-health danger of using too large a block size?* •

A10 : *The larger the block, the more the averaging effect. So a local polluted hot spot may get averaged with unpolluted areas, so that the block average is below the cleanup threshold, and the entire block is not treated. The hot spot remains dangerous.* *Return to Q10* •

Answers

Q11 : *Considering the long-range variogram, what spacing between observations would be needed to obtain the same or lower kriging variance for a 20 m block as using a 400 m spacing and a 120 m block?* •

A11 : *The kriging variance for a 400 m spacing and a 120 m block is about 0.6 (dark purple colour in the map); this is reached for the 20 m block at a 250 m spacing.* *Return to Q11* •

Q12 : *If now the variogram is short-range (i.e., the spatial dependence is only over shorter separations), what is the widest spacing that could be used to obtain the same or lower kriging variance as using a 400 m spacing and a 120 m block for the long-range variogram?* •

A12 : *Looking for 0.6 (dark purple) along the spacing axis of the short-range variogram, we see that 250 m spacing for the largest block size (120 m) would give this variance. So the reduction in variogram range from 1200 to 600 m (i.e., half) results in a denser sample spacing to obtain the same maximum kriging variance.* *Return to Q12* •

Answers

Q13 : *Why should the procedure ever accept a* **worse** *scheme?* •

A13 : *To avoid getting stuck in a* **local minimum**; *sometimes a point should "jump" a long way. Return to Q13* •