Sampling

Justification

How many observations *where* to take
- Taking data is expensive
- Many data are needed to characterize an area
- Optimal sampling requires a definition of optimality.

Relevant in:
- Agricultural modeling
- Remote sensing (GCPs)
- Environmental sampling (air quality, water sediments)
- Animal tracking
- Soil ecology

Survey vs. design

- Empirical research: surveys, experiments
  - Surveys: investigators observe
  - Experiments: investigators actively influence and manipulate
- Need for efficient (cost-effective) research, optimal research designs
- Therefore need for detailed planning *before* experiment is performed

Part 1

Experimental design

Experimental design

- Statistics:
  - Data gathering, e.g. experimental design
  - Data analysis; descriptive statistics vs inferential statistics
  - Drawing conclusions
- Statistical Power
  - Introduction
  - Estimating power

Empirical research

- Formulation of the problem
- Definition of precision requirements
- Selecting the statistical model for planning and analysis
- The (optional) design of the experiment or survey
- Performing the experiment or survey
- Statistical analysis of the observed results
- Interpretation of the results

Definitions

- The population is a set of objects or individuals having certain characteristics which we wish to make a statement about
- An experiment is
  - the active influence of research workers with the aim of observing the result of this influence
  - the observation of one or of several elements of a well defined population or universe

Definitions

- An experimental unit is the individual or object which is treated in the experiment.
- The sampling unit is an element of the population which as a result of the sampling process is included in the sample

Planning of Experiments and Surveys and the Description of simple Designs

- Basic Ideas
  - Formulate the research question as precisely as possible
  - Choose the appropriate statistical model
- Precision requirements for results
- If an optimal design is needed, optimality criterion

Planning of Experiments and Surveys and the Description of simple Designs

- Three R's in planning of experiments
  - **R**eplication
  - **R**andomization
  - **R**educing noise factors

Planning of Experiments and Surveys and the Description of simple Designs

- Principle of **R**eplication:
  - Let y have a distribution with expectation μ and variance σ^2
  - Let $y_1,..,y_n$ be independent and distributed as y
  - Then average(y) is a random variable with expectation μ and variance σ^2 /n
- Replication decreases the variability of the sample average
- To estimate of variance needed: n > 1
- Replication may secure the desired precision e.g. average(y) must fall in interval (μ-σ/8, μ+σ/8) with prescribed probability; probability is a function of n

Planning of Experiments and Surveys and the Description of simple Designs

Use Strata or Blocks to eliminate the Effects of Noise factors

- Eliminate noise factors as much as possible e.g. by
  - standardisation
  - analysis of covariance
  - forming blocks or strata
  - randomization
- Terminology
  - factors, factor levels
  - design factors: treatment factors and block factors
  - noise factors: block factors and residual

Planning of Experiments and Surveys and the Description of simple Designs

- Randomization / random sampling: appropriate if noise factors unknown
- Blocking, stratification: grouping together experimental units with same level of known or suspected noise factor

Terminology

+ (statistical) experimental design
+ statistical planning of experiments: construct design, determine minimum sample size
+ statistical model
+ optimal design
+ treatment=level of treatment factor or combination of levels of treatment factors
+ factorial design
+ block design, blocks, block size
+ complete, incomplete block design
+ balanced incomplete block design

Planning of Experiments and Surveys and the Description of simple Designs

Randomization is a procedure to randomly select EUs from a population in a survey and to randomly assign EUs to treatments in an experiment

Randomization in Surveys - Random Sampling

- Random sampling: all N elements of population have same chance of being chosen
- Unrestricted random sampling: all "N over n" possible subsets have same probability.
- Random sampling in practice, e.g. use random number generator, systematic sampling
- Stratified sampling

Randomization in Experimental Designs - Random Allocation

- Unrestricted randomization - completely randomized design
- Restricted randomization for block design

Part 2

Design based sampling

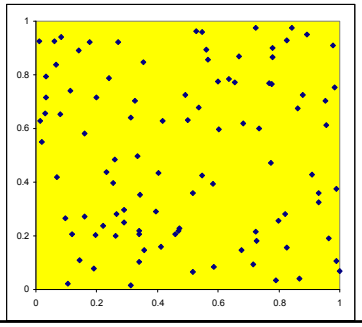
Design-based sampling

- Basic idea: each point in the area has the same probability of being sampled
- Sampling points are assigned to a location in an area (space or space/time)
- This is done in a random way
- Examples:
  - Complete random sampling
  - Stratified random sampling
  - Grid sampling
  - Cluster sampling

Complete random sampling

- The area is available as a polygon
- The number *n* of sampling points is decided beforehand
- j <- 1
- do j < n
  - Using a random number generator, the *x*- and *y*-coordinates of a sampling points are drawn
  - If the point occurs inside the area then the point is accepted and j <- j+1

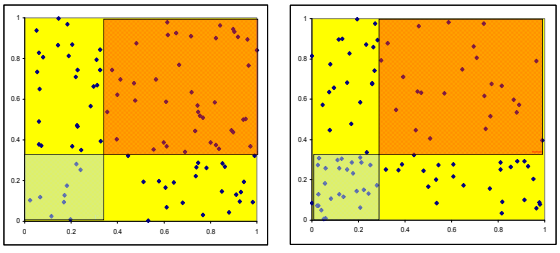
Complete random sampling



Stratified random sampling

- The area is stratified on the basis of available information into *k* strata A_i
  - A soil map
  - Segments from a segmented RS image
  - Available boundaries
- The points are distributed: n_1 for A_1, n_2 for A_2, etc.
- Distribution could be according to size of the stratum, might include a minimum number of points, equally many points in each stratum, etc.
- SRS is applied within each stratum

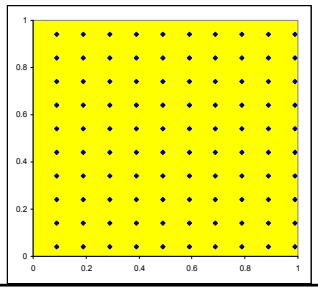
Stratified random sampling



Grid sampling

- The area is available as a polygon
- The number *n* of sampling points is decided beforehand
- A regular grid is defined in the area, containing *n* grid points
- The grid is moved over a random vector so that all grid points are falling within the area
- Grid sampling can be combined with stratification
- Often a pragmatic solution is necessary to deal with irregular boundaries

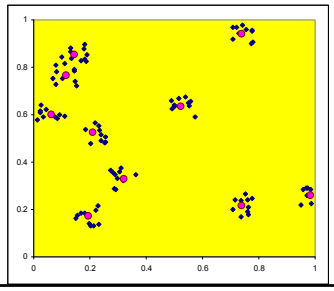
Grid sampling



Cluster sampling

- In cluster sampling we apply a 2-layer approach
- The area is available as a polygon
- The number *n* of sampling points is decided beforehand
- A low number (n_1) virtual points are randomly assigned in the area
- Around each of these, usually at a small distance, n/n_1 points are assigned and are actually sampled
- This is relatively cheep sampling

Cluster sampling



Part 3

Model based sampling

Model based sampling

A random model is at the basis of the design, for example a random field model

There is a quantitative optimization criterion

- Triangular scheme is optimal for infinite populations
- Prior data and irregular boundaries can easily be considered

Each criterion leads to a unique optimal scheme

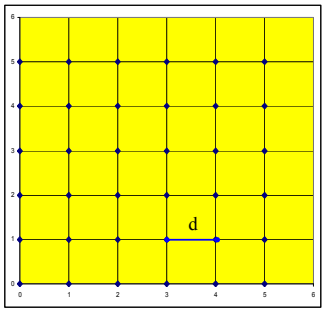
The kriging variance

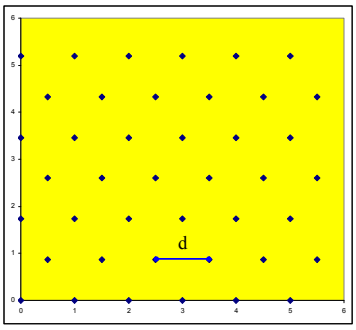
$$Var(T - Z(x_0)) = c_{00} - c_0^T C^{-1} c_0 + x_a^2 / V$$

where $x_a = 1 - c_0^T C^{-1} 1^n$
and $V = (1_n^T C^{-1} 1_n)$
Kriging standard error = sqrt(Var)

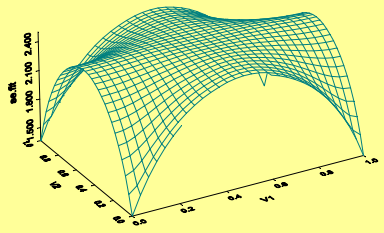
Starting

- Realize that the kriging variance does not depend upon data values
- Suppose that a variogram is available
- Then the data configuration fully determines the final precision
- In the literature: triangular, square and hexagonal schemes
- Largest uncertainty in center points

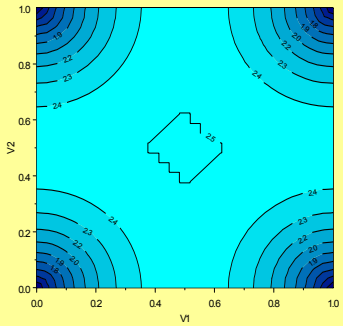




Kriging Std. Errors



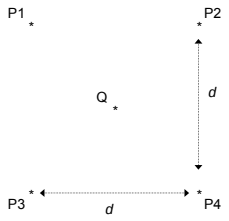
Kriging Std. Errors



Example

- Take a square grid with grid mesh equal to d.
- Predictions with highest prediction error variances occur in the centre point of the grid cells
- Assume that the variogram is a linear model without sill, g(**h**) = |**h**|.

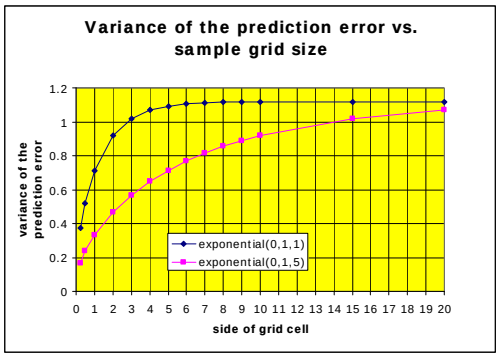
Consider the following configuration of the data points P_1 t/m P_4 and determine the prediction error variance in the point Q:



With a linear variogram, the standard deviation of the prediction error increases proportional to the square root of the grid mesh.

If we want to obtain predictions with the prediction error variance equal to 1, the grid mesh should be equal to d = 1/0.6 = 1.67.

This procedure gives the grid mesh by given variogram and a given preset precision.



Problems in spatial sampling

- A regular grid (triangular, square) is usually recommended for optimal sampling
- Objection: a grid aims at minimising *Euclidian* distance to the nearest observation. We wish to include as well *other* optimisation criteria

Optimising

- Optimisation function:

  Minimisation of the mean distance from an arbitrary point in the region towards the closest observation point

- Approximated by the *fitness function*:

$$\varphi = \min_{x_j \in S} \frac{1}{n_e} \sum_{i=1}^{n_e} \min_j d(\tilde{x}_i, x_j)$$

Different criteria

- Equal numbers of pairs of points in distance classes for variograms.
- Equal spreading of points in the space-time domain
- Lowest maximum kriging variance (= the best map)
- Lowest average kriging variance
- Best classification (or segmentation)
- Any other relevant quantitative criterion

Take into account

- Prior data and information
- Spatial variability
- Boundaries and restrictions
- Subjective weights of importance for subdomains

Spatial Simulated Annealing

Step 1:

- Random sampling scheme S_1
- Calculate fitness function

Step 2:

- Generate S_2 using a transformation of a random sampling point over a random vector
- Calculate fitness function for S_2

- Step 3:

  Accept of S_2 if (Metropolis criterion):

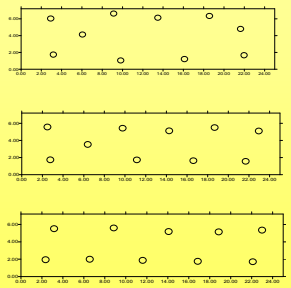
$$P_c(S_1 \to S_2) = 1, \qquad \text{if } \phi(S_2) \le \phi(S_1)$$

$$P_c(S_1 \to S_2) = \exp\left(\frac{\phi(S_1) - \phi(S_2)}{c}\right), \qquad \text{if } \phi(S_2) > \phi(S_1)$$

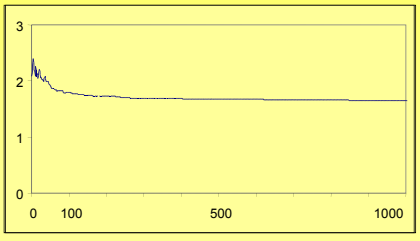
- Step 4:
  - The process returns to Step 2, where S_3 will be generated out of S_2 or S_1.

Criterion 1: Even distribution of points

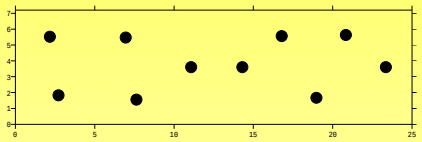
- Three intermediate steps in optimising the sampling scheme - after 100, 500 and 1000 iterations

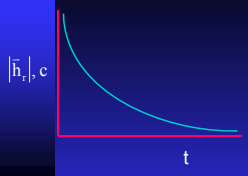


The fitness function during optimization



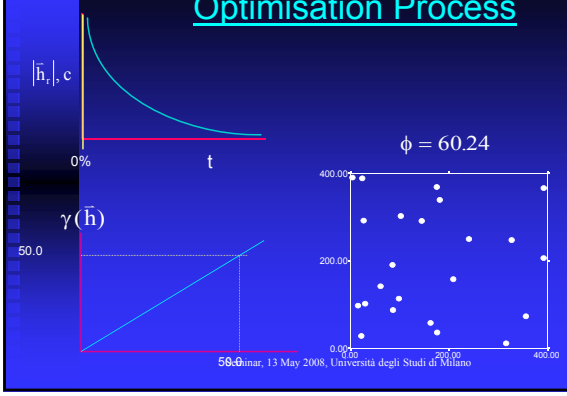
Optimization with 3 preference points

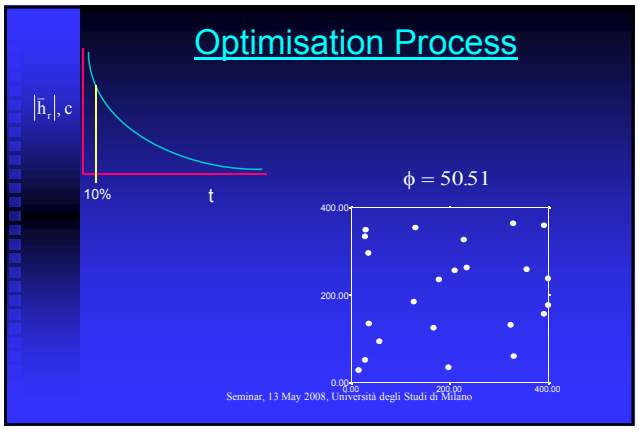


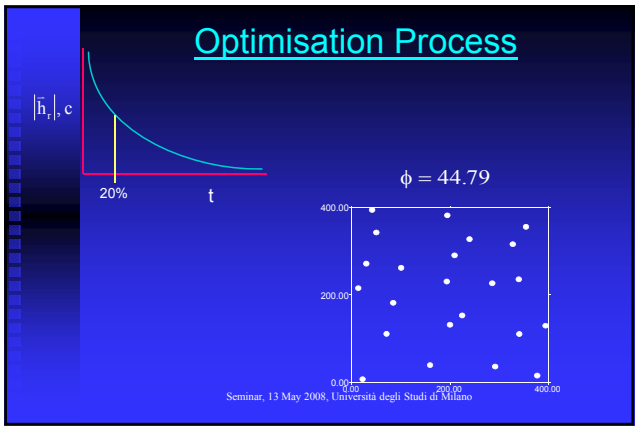


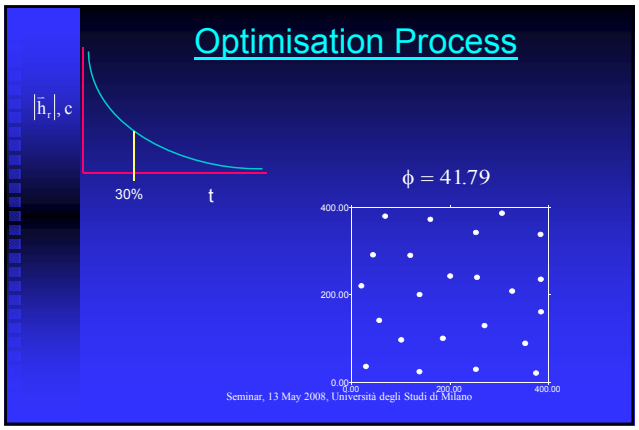
To ensure convergence towards minimum, *c* and the maximum of $|\vec{h}_r|$ are lowered during optimisation

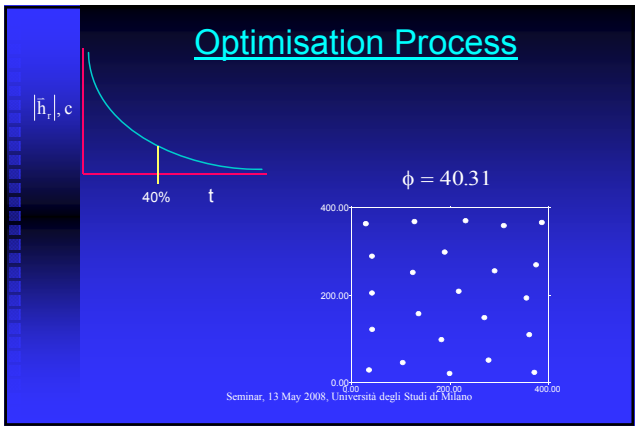
Optimisation Process

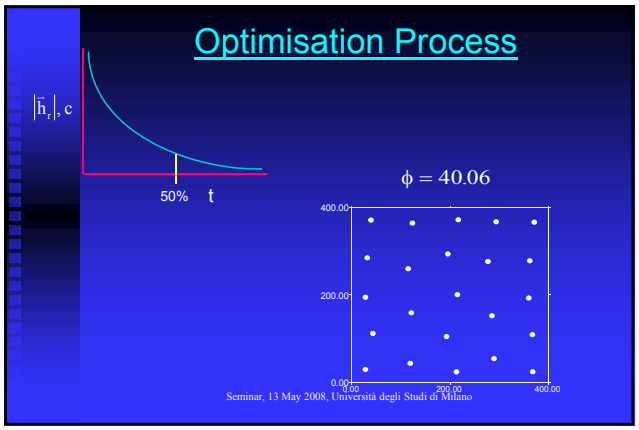


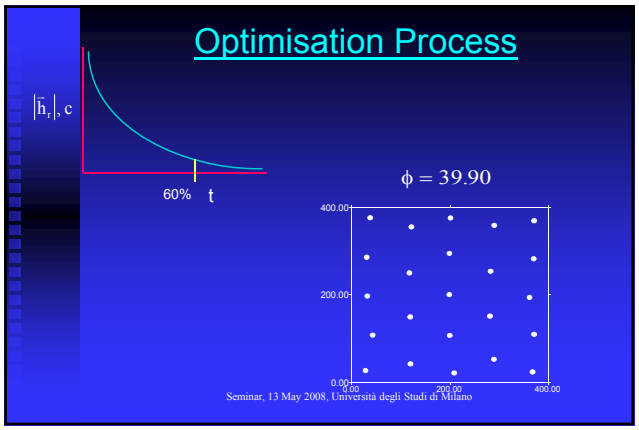

Optimisation Process
$|\bar{h}_r|, c$
10%
t
$\phi = 50.51$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano

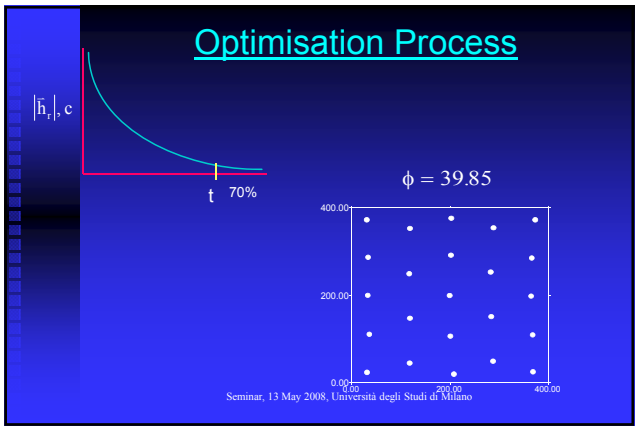

Optimisation Process
$|\bar{h}_r|, c$
20%
t
$\phi = 44.79$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano

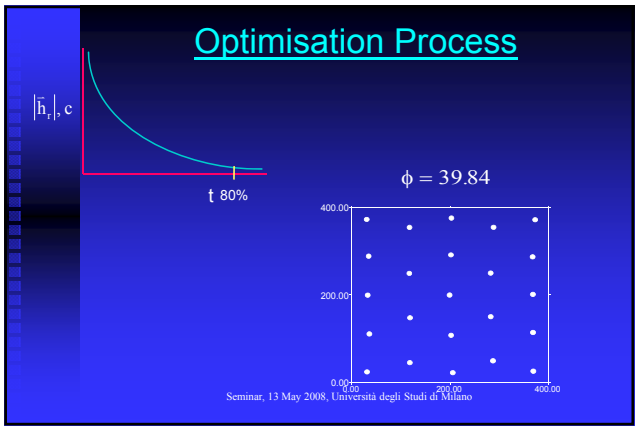

Optimisation Process
$|\bar{h}_r|, c$
30%
t
$\phi = 41.79$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano

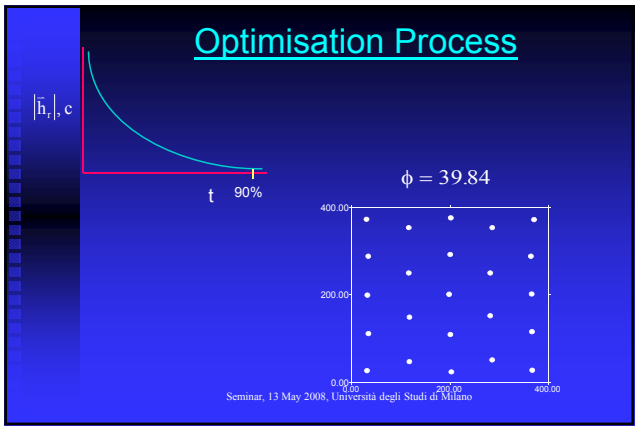

Optimisation Process
$|\bar{h}_r|$, c
40%
t
$\phi = 40.31$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano

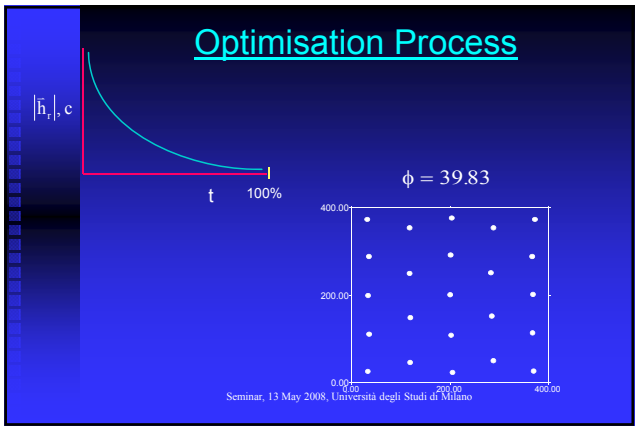

Optimisation Process
$|\bar{h}_r|$, c
50%
t
$\phi = 40.06$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano


Optimisation Process
$|\bar{h}_r|$, c
60%
t
$\phi = 39.90$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano


Optimisation Process
$|\bar{h}_r|$, c
t 70%
$\phi = 39.85$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano

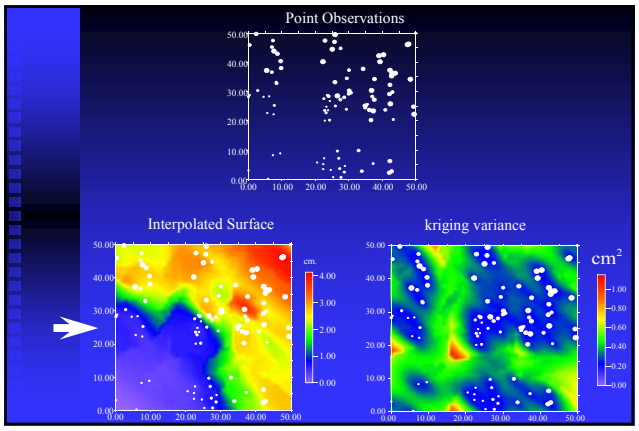

Optimisation Process
$|\bar{h}_r|$, c
t 80%
$\phi = 39.84$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano


Optimisation Process
$|\bar{h}_r|$, c
t 90%
$\phi = 39.84$
400.00
200.00
0.00
0.00
200.00
400.00
Seminar, 13 May 2008, Università degli Studi di Milano



Criterion 2: the best possible map

- To design sampling schemes with minimal kriging variance
- To show the effects of kriging parameters on the optimal sampling scheme



Optimising

- Minimisation function:

$$\int \sigma^2_{OK}(\bar{x}) \cdot d\bar{x}$$

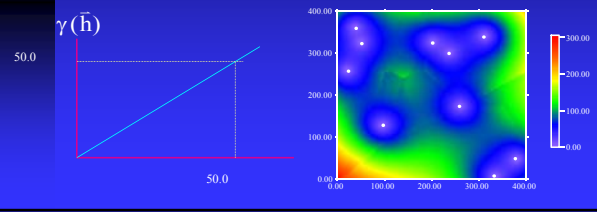
- Approximated by the *fitness function*:

$$\phi(S) = \sum_{j=1}^{n_e} \frac{\sigma^2_{OK}(\bar{x}_{e,j})}{n_e}$$

Spatial Simulated Annealing

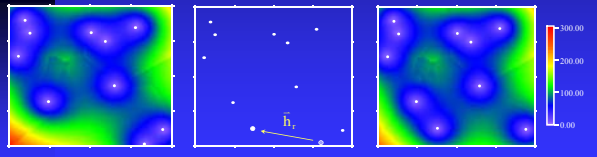
Step 1:

- Variogram (linear)
- Random sampling scheme S_1 (10 points)
- Calculate fitness function $\phi(S_1) = 86.87$



- Step 2:
  - Generate S_2 using a transformation of a random sampling point over a random vector
  - Calculate fitness function for S_2

$\phi(S_1) = 86.87$ $\phi(S_2) = 80.47$

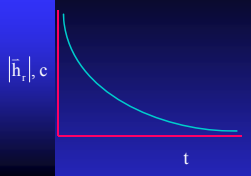


- Step 3:
  - acceptance of S_2 depends on Metropolis criterion:

$$P_c(S_1 \rightarrow S_2) = 1, \qquad \text{if } \phi(S_2) \leq \phi(S_1)$$

$$P_c(S_1 \rightarrow S_2) = \exp\left(\frac{\phi(S_1) - \phi(S_2)}{c}\right), \qquad \text{if } \phi(S_2) > \phi(S_1)$$

- Step 4:
  - The process returns to Step 2, where S_3 will be generated out of S_2 or S_1.



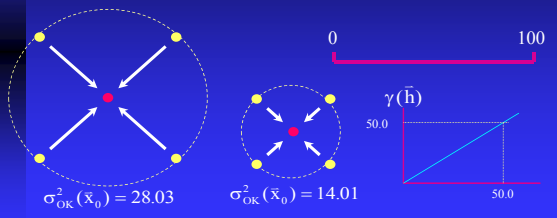
To ensure convergence towards minimum, *c* and the maximum of $|\bar{h}_r|$ are lowered during optimisation

KRIGING VARIANCE DEPENDS ON:

- variogram
- neighbourhood
- sampling scheme

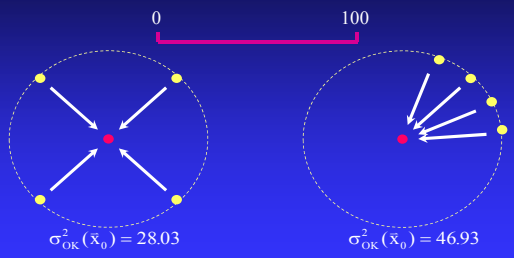
KRIGING VARIANCE DEPENDS ON:

- Distance between sampling points and estimated point:



KRIGING VARIANCE DEPENDS ON:

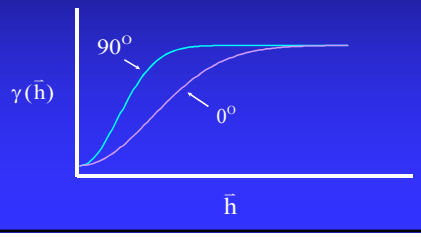
- Distance between sampling points :

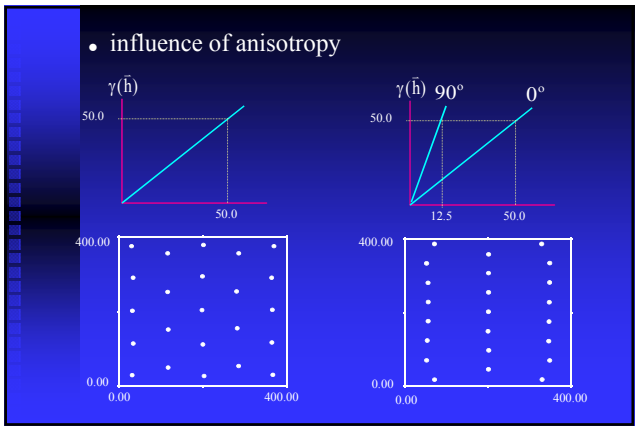


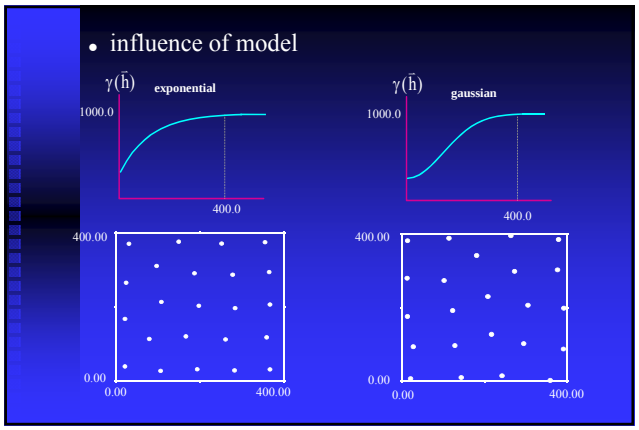
KRIGING VARIANCE DEPENDS ON:

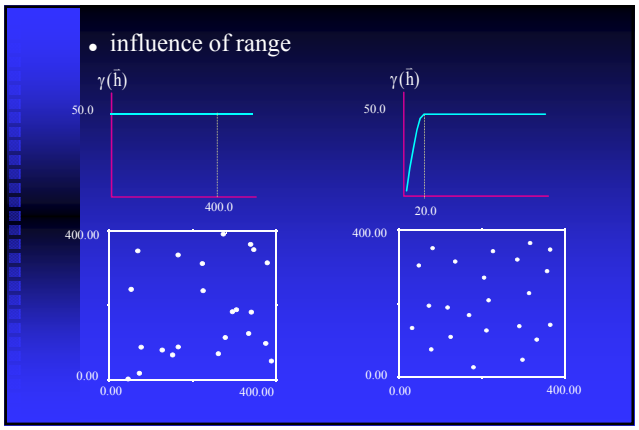
- Variogram shows spatial correlation:

$$\hat{\gamma}(\vec{h}) = \frac{1}{2n(\vec{h})} \sum_{i=1}^{n(\vec{h})} \{z(\vec{x}_i) - z(\vec{x}_i + \vec{h})\}^2$$




influence of anisotropy
γ(h)
50.0
50.0
γ(h)
90°
0°
50.0
12.5
50.0
400.00
0.00
0.00
400.00
400.00
0.00
0.00
400.00

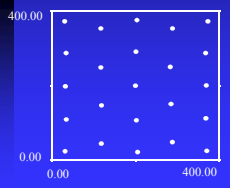

influence of model
γ(h)
exponential
1000.0
400.0
γ(h)
gaussian
1000.0
400.0
400.00
0.00
0.00
400.00
400.00
0.00
0.00
400.00


influence of range
γ(h)
50.0
400.0
γ(h)
50.0
20.0
400.00
0.00
0.00
400.00
400.00
0.00
0.00
400.00

Minimising mean vs. max

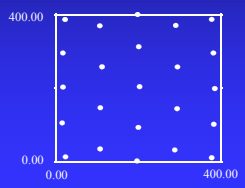
Minimising mean Kriging variance:

$$\phi(S) = \sum_{j=1}^{n_e} \frac{\sigma^2_{OK}(\bar{x}_{e,j})}{n_e}$$



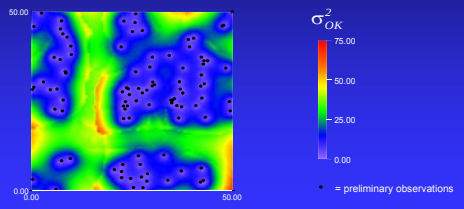
Minimising max Kriging variance:

$$\phi(S) = \max(\sigma^2_{OK}(\bar{x}_{e,j}), \forall\, \bar{x}_{e,j})$$



Minimising mean vs. max

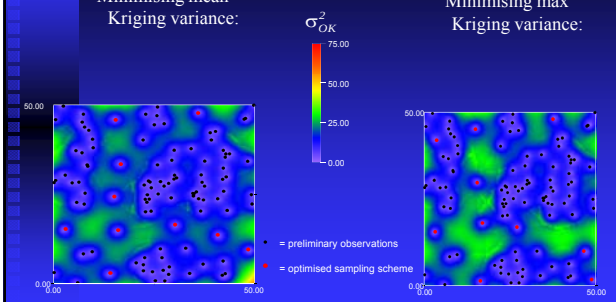
Preliminary sampling scheme: 100 irregular observations:



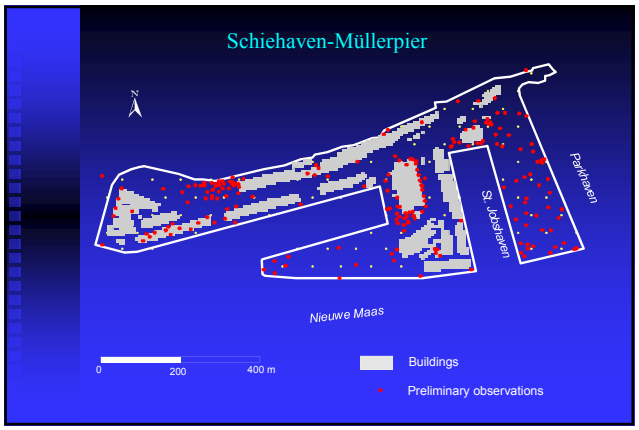
Minimising mean vs. max

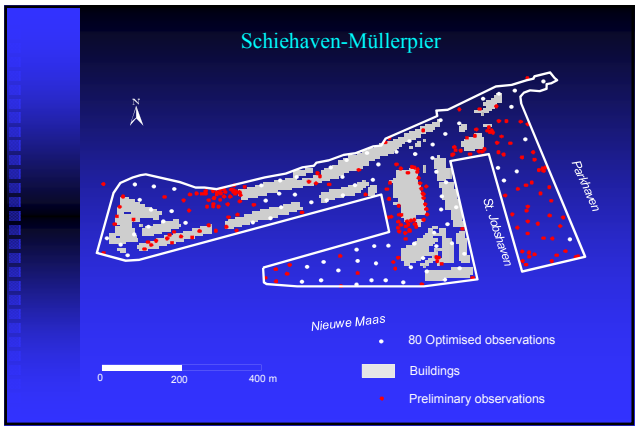
Minimising mean Kriging variance:

Minimising max Kriging variance:




Schiehaven-Müllerpier


Schiehaven-Müllerpier
N
Parkhaven
St. Jobshaven
Nieuwe Maas
0
200
400 m
Buildings
Preliminary observations


Schiehaven-Müllerpier
N
Parkhaven
St. Jobshaven
Nieuwe Maas
0
200
400 m
80 Optimised observations
Buildings
Preliminary observations

Sampling of fuzzy areas

Sampling

- Sampling identifies of objects in space.
- Optimal sampling relies on the selection of an objective function that we aim to optimize with an appropriate sampling design
- For vague objects
- 1. Characterize the objects that we can be sure about
- 2. Characterize the transition zones, i.e. those parts of an area that we are least sure about

The confusion index

- Fuzzy classification depends on the number of classes n and the fuzziness index φ.
- It yields membership values for all classes
- Final classification can be done by selecting at any point the class with the highest membership value
- The ratio of the 2nd and 1st largest membership value is the confusion index
- After classification, an area can be split it a subarea A_C with a low confusion index and a subarea A_F with a high confusion index.

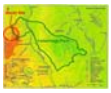
Objective function

Optimal sampling here optimizes sampling according to the following objective function

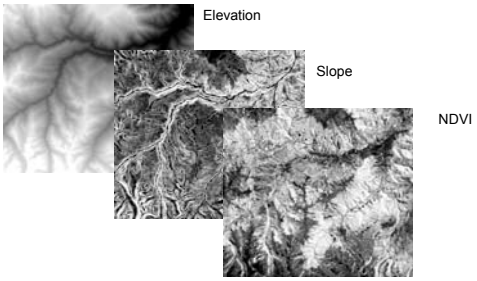
$$\phi(S) = \int_{A_C} \|x - V_C(x)\|^2 dx + \int_{A_F} \|x - V_F(x)\|^2 dx$$

where x is a location vector
$V_C(x)$ is the location vector of the nearest sampling point $x_i \in S|A_C$
$V_F(x)$ is the location vector of the nearest sampling point $x_i \in S|A_F$.
We aim to determine S^* which minimizes $\phi(S)$.
Starting with CRS, derive S^* with simulated annealing

Study area:Yanachaga Park, Peru

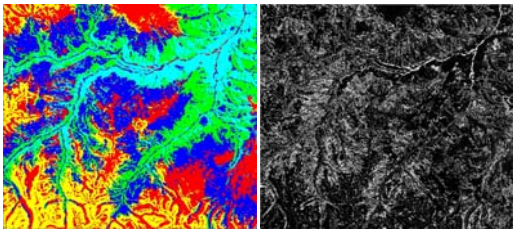


Additional information

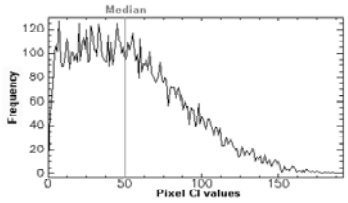


Fuzzy classification

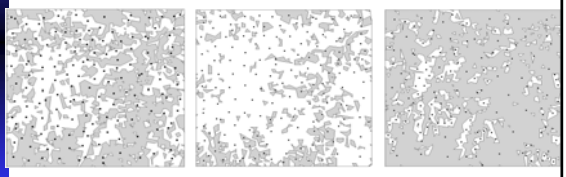
A fuzzy classification with φ = 1.5 into 5 classes (left) and the corresponding confusion index (right)



Histogram of pixels on the confusion index map



Allocation of 150 sample locations over the study area: 100 samples are located in the low confusion area, and 50 samples in the high confusion area



Median of CI values | $Q_{0.25}$ of CI values | $Q_{0.75}$ of CI values