Model-based vs. design-based sampling strategies for monitoring, with a case study on testing surface water quality against standards

Martin Knotters, Dick Brus

Soil Science Centre

Wageningen University and Research Centre - Alterra

Aim

- To discuss two basically different sampling strategies: model-based and design-based
  - What are the basic differences?
  - How to choose?
- To illustrate the application of a design-based (probability-based) sampling strategy in compliance monitoring of surface water quality

Motivation

- Selection process of sampling sites and moments deserves more attention: 'representative' is often ill-defined*
- 'Start at the end, and reason backward': integrated planning of data collection and data processing, with respect to the required information
- Probability sampling can be useful, but is seldom applied in WFD monitoring (in contrast to USA, Clean Water Act)

*) Eight different meanings. See for an interesting analysis of the various meanings of 'representative sampling': Kruskal, W. and F. Mosteller, 1979. Representative sampling, I, II and III. *International Statistical Review* **47**(1,2,3): 13-24, 111-127, 245-265.

Design-based and model-based approach

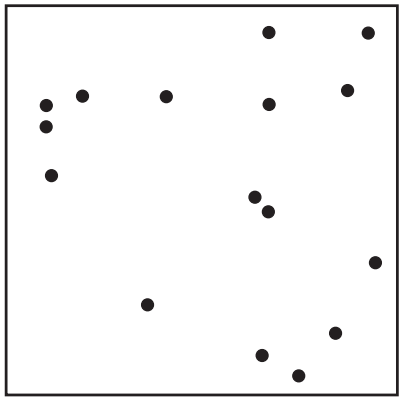
Definition of design-based and model-based approach

Type of approach	Sampling unit selection	Statistical inference
Design-based	Probability sampling	Design-based
Model-based	No requirement (purposive)	Model-based

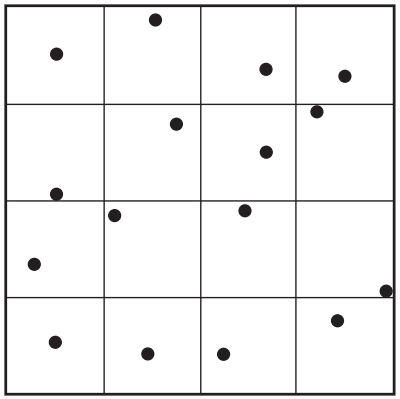
Probability sampling

- *random* selection of elements
- selection probabilities are known, and > 0
- inference based on selection probabilities

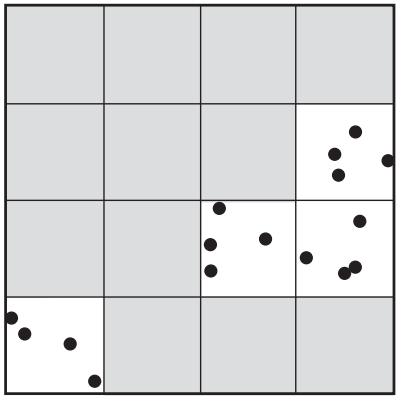
Sampling design types



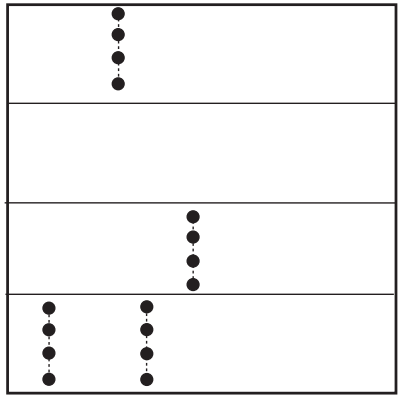
Simple random sampling



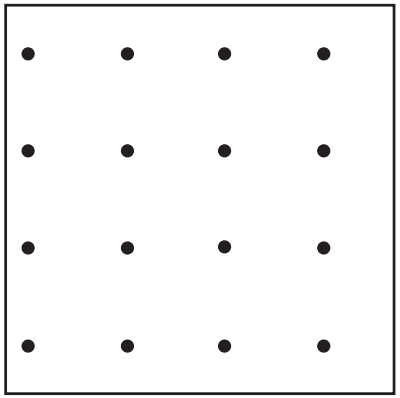
Stratified simple random sampling



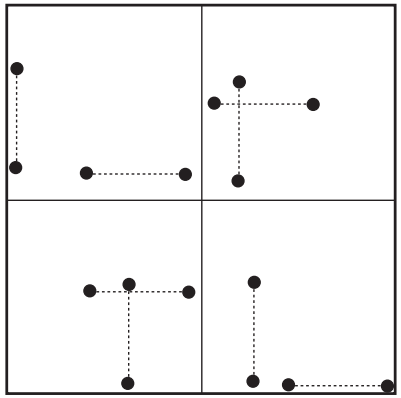
Two-stage random sampling



Cluster random sampling



Systematic random sampling



Stratified cluster random sampling

Design-based or model-based approach?

Design-based method best choice when:

- we want to estimate the distribution function or parameters thereof (mean, median, P90 etc.) for the area as a whole or for some subareas;
- a minimum sample size of 10 units (per subarea) can be afforded;
- valid results are required, i.e. the quality is independent of the quality of model assumptions: For example: testing against standards, validation studies;
- it is practically feasible to sample at randomly selected locations.

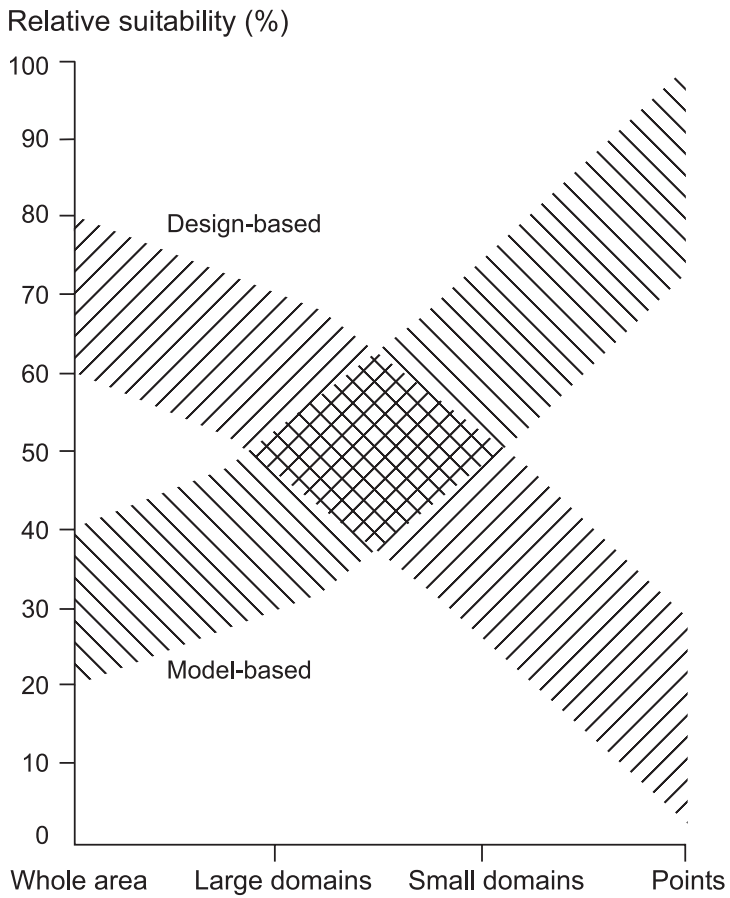
Design-based or model-based approach?, continued

Model based method best choice when:

- we want to map the target property;
- sample size large enough for calibrating a model of variation (e.g. variogram: $n > 100$);
- strong autocorrelation exists, from which we may profit in mapping;
- system analyses, scenario studies, etc..

Further reading: D.J. Brus and J.J. de Gruijter (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). Geoderma, 80: 1-59.

Suitability statistical approaches

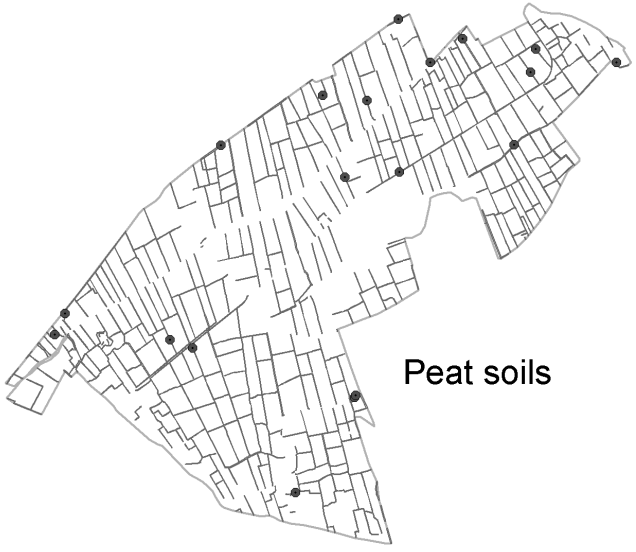


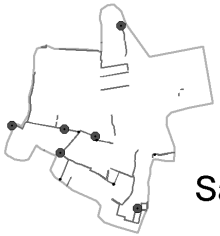
Case study: testing surface water quality against standards

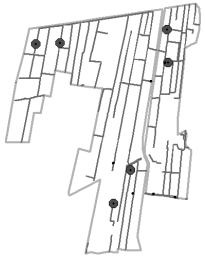
Northern Frisian Woodlands: does space-time mean concentration total N in surface water during summer half-year comply with standard of 2.2 mg/l?

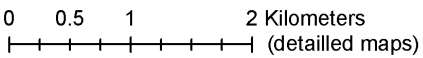
Study area


Clay soils

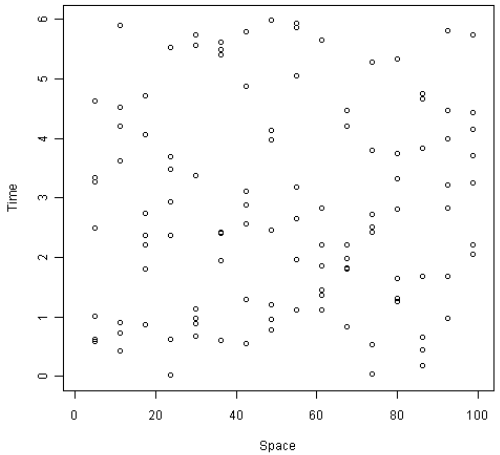

Peat soils


Sandy soils

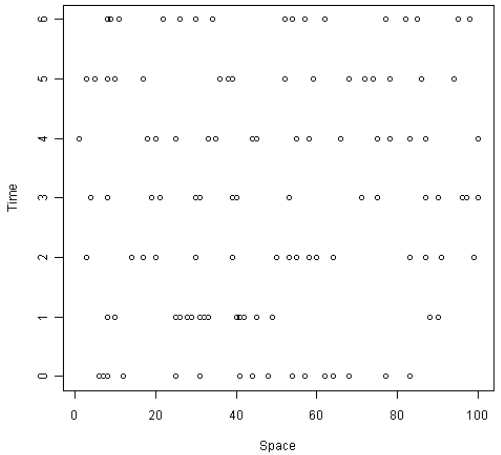

Sandy and peaty soils



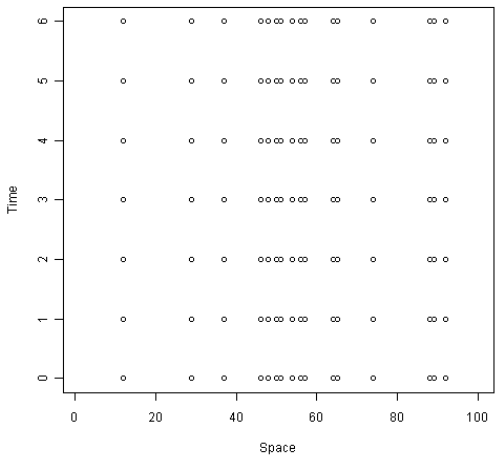
Sampling in space and time: pattern types



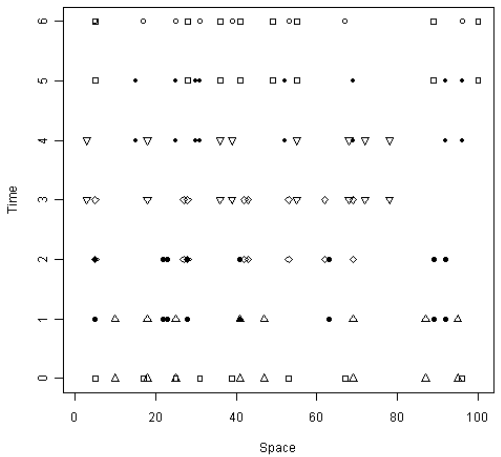
Static



Synchronous

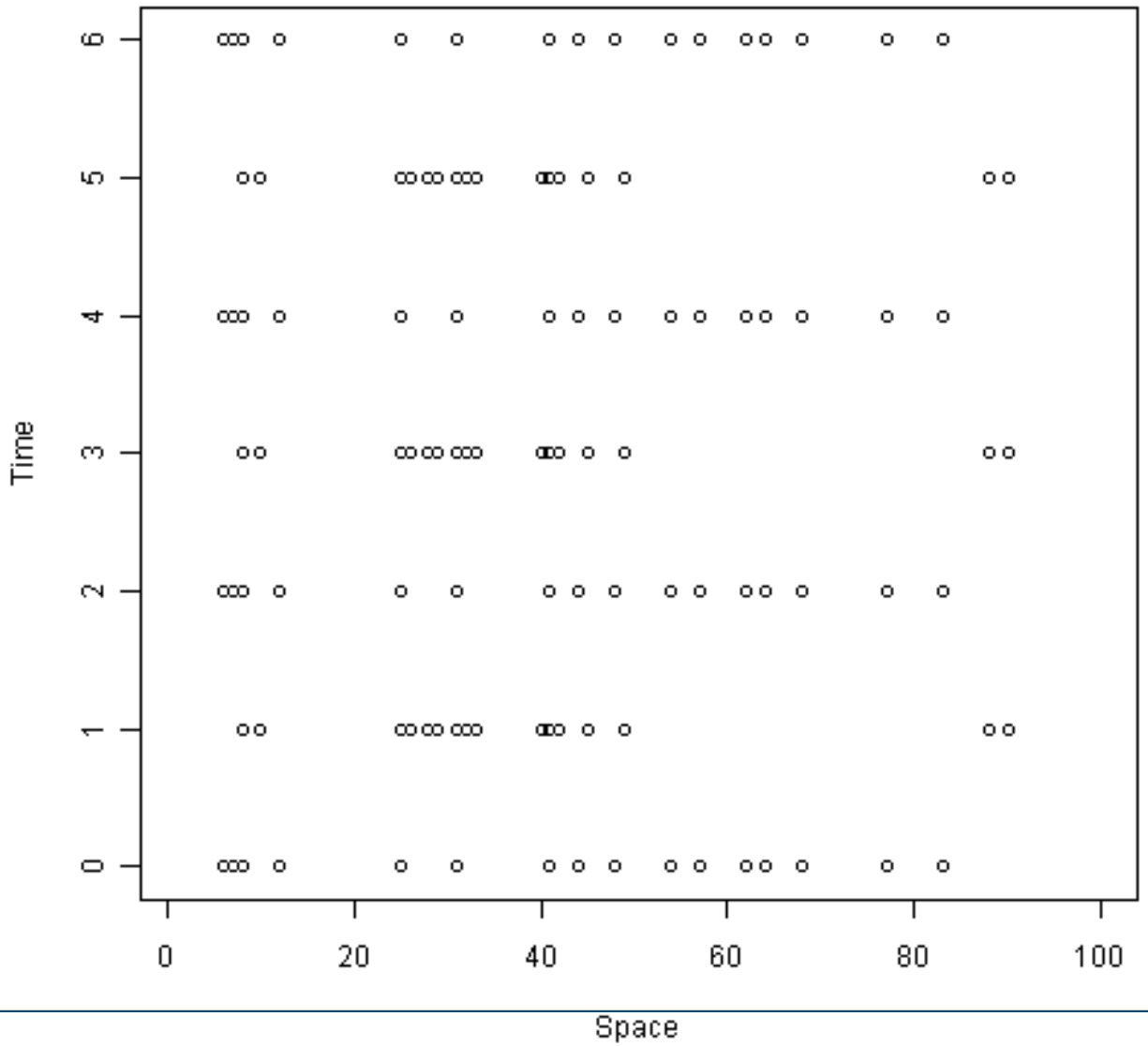


Static-Synchronous

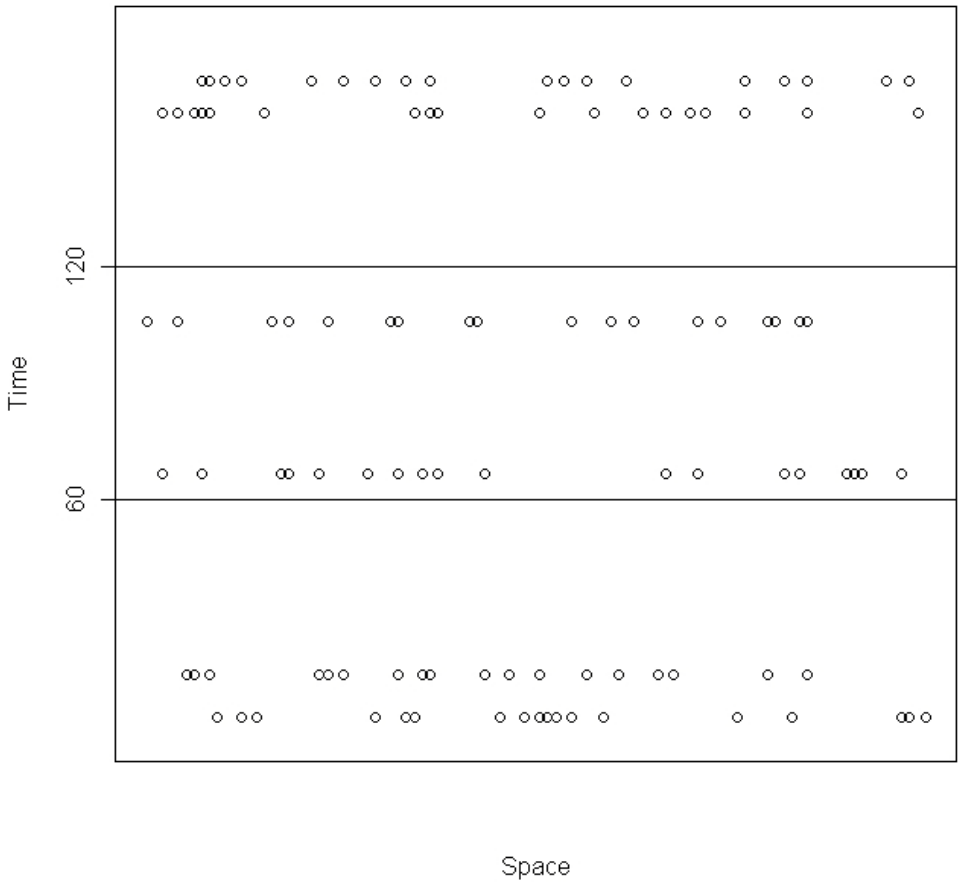


Rotational

Serially alternating



Synchronous sampling pattern



Stratified simple random sampling in time (STSI), simple random sampling in space (SI)

Space-time mean concentration Total N

$$\hat{\bar{y}}_{\mathrm{STSI,SI}} = \frac{\sum_{h=1}^{\ell} \frac{T_h}{n_h} \sum_{i=1}^{n_h} \frac{A}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \cdot l_{hij}}{\sum_{h=1}^{\ell} \frac{T_h}{n_h} \sum_{i=1}^{n_h} \frac{A}{m_{hi}} \sum_{j=1}^{m_{hi}} l_{hij}}$$

with

y_{hij}: concentration measured at location j, time i in temporal stratum h

l_{hij}: measured water depth

A: (surface) area of study area

T_h: length of temporal stratum

Results

Estimated space–time mean concentrations of Total N in surface water, in four subareas of the Northern Frisian Woodlands, in the period April 1, 2008 to September 30, 2008.

Subarea	$\hat{\bar{y}}_{\mathrm{STSI,SI}}$ (mg·l^{-1})	$\mathrm{var}(\hat{\bar{y}}_{\mathrm{STSI,SI}})$ mg^2· l^{-2}
1 (peat soils)	2.45	0.017
2 (sandy soils)	3.44	0.493
3 (sandy and peaty soils)	2.23	0.066
4 (clay soils)	2.31	0.030

(Standard: 2.2 mg/l)

Results (continued)

Optimized number of sampling rounds n and locations per round m, for a budget that equals the costs in 2008

Subarea	Summer 2008		Optimized	
	n	m	n	m
1 (peat soils)	6	18	16	5
2 (sandy soils)	6	6	5	7
3 (sandy and peaty soils)	6	6	3	14
4 (clay soils)	6	10	7	8

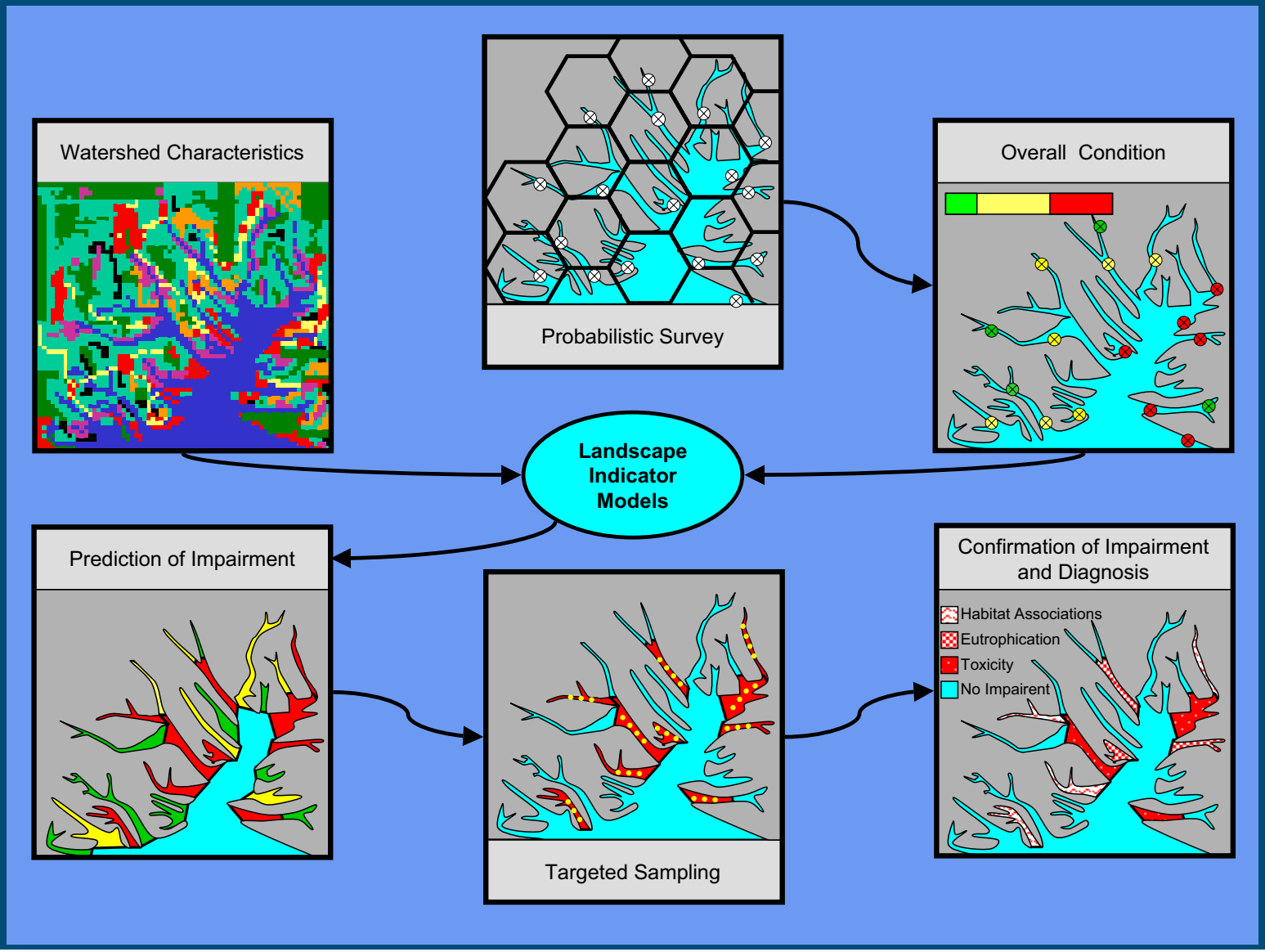
Further reading

De Gruijter, J.J., D.J. Brus, M.F.P. Bierkens & M. Knotters, 2006. *Sampling for natural resource monitoring*. Berlin, Springer, 332 pp.

Brus, D.J. & M. Knotters, 2008. Sampling design for compliance monitoring of surface water quality: A case study in a Polder area. *Water Resources Research* **44**, W11410.

Knotters, M. & D.J. Brus, 2010. Estimating space-time mean concentrations of nutrients in surface-waters of variable depth. *Water Resources Research*, in press.

Probability-based sampling for Clean Water Act (EMAP, USA; Nicholson, 2004)



Probability-based sampling for Clean Water Act (EMAP, USA)

http://www.epa.gov/owow/monitoring/
http://www.epa.gov/emap/index.html
http://www.epa.gov/305b/

Concluding remarks

- Is modelling inevitable in monitoring? No!
- Design-based methods attractive for WFD monitoring: global information, testing against standards, model validation

Thank you!