Applied geostatistics

Lecture 7 – Geostatistical risk mapping

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

January 7, 2014

Topics for this lecture

1. Uncertainty, hazard and risk

2. Indicator variables

3. Indicator variograms

4. Probability kriging with indicator variables

Topic 1: Uncertainty, hazard and risk

These three terms are increasing order of difficulty to evaluate.

1. In geostatistics we always attempt to quantify **uncertainty**;
2. This can be converted to a **hazard** assessment, given appropriate thresholds
3. The final **risk assessment** is usually then left to other specialists.

We now define each of these terms.

Uncertainty

Uncertainty: lack of knowledge about the true state of nature

- "What is the concentration of cadmium in the shallowest 20 cm of 10x10 m area of soil centred at UTM coördinates ..."
- Preferably **quantified** as the **probability** of any state
- "Lognormally-distributed with expected value 3, standard deviation 1"

Hazard

Hazard: the chance of a given (bad) condition or outcome

- "How likely is it that the Cd concentration exceeds the regulatory threshold of 2 mg kg^{-1}?"
- Expressed probabilistically: "$p < 0.02$ that rejecting the null hypothesis of no risk would be an error", i.e. it's most likely polluted!

Risk

Risk: the chance of something 'bad' happening

- "How likely is it that children playing soccer on a grass field at location UTM . . . will become poisened by cadmium?"
- Must be **quantified** as the **probability** of the outcome
- Requires full specification of **elements at risk**, **exposure** and **vulnerability**

Assessing uncertainty

- This requires the specification of a **probability distribution** of a predicted value
- This can be provided by the **kriging prediction variance**

Assessing hazard

- Method 1: Convert the **uncertainty** (full probability distribution) to a single probability of exceeding a **threshold** value
  - E.g. confidence intervals from kriging prediction and its variance
  - This was shown in a previous lecture
  - Requires **strong assumptions** about the probability distribution of the target variable
- Method 2: Predict the **probability of occurrence** directly, using an **non-parametric** methods.

We will continue with the second method (below).

Assessing risk

- Knowing the **hazard**, determine the **exposure pathway** from the given hazard to a defined **risk**.

- The risk must be defined from the characteristics of the **element at risk**
  - e.g., human health: what concentrations and exposure times / methods lead to a given condition?
  - this is well outside geostatistics, it requires extensive **domain knowledge**

We continue with the **uncertainty** estimate.

Distribution-free estimates

So far we have assumed an approximately normal or lognormal distribution of the target spatially-correlated random variable. But this may be demonstrably not true.

A **non-parametric** approach does not attempt to fit a distribution to the data, but rather works directly with the experimental CDF, by dividing it into sample **quantiles**.

To work with these, we introduce the idea of **indicator** variables.

Note: there are other methods, such as **disjunctive kriging**, which we do not cover here.

Note: the indicator kriging approach has been strongly criticized on theoretical grounds, see e.g., Papritz (2009) in the "Further reading" (below).

Topic 2 : Indicator variables

- **Binary** variables: Take one of the values $\{1, 0\}$ depending on whether the point is 'in' or 'out' of the set; i.e. if it does or does not meet some criterion
  - ∗ These are suitable for binary **nominal** variables, e.g. {"urban", "not urban"}; {"land use changed", "land use did not change"}

- A **continuous** variable can be converted to an indicator z_t by a **threshold** or **cut-off** value x_t: $z_t = 1 \iff x \leq x_t$
  - ∗ e.g. $x_t = 350$ to cut-off at 350 mg kg^{-1}
  - ∗ Formally: $I(\vec{x}_i, z_t) = 1$ iff $Z(\vec{x}_i) \leq z_t$; 0 otherwise
  - ∗ By **convention** 1 indicates values **below** the threshold (to model the CDF); inverting reverses the sense
  - ∗ Note we are **losing all information** from the continuous variable, **except** at the **cut-off**!

Computation of indicators

Here are the first 16 observations in the Meuse soil pollution data set, with their actual Cd values and an indication of whether they are below a thresholds of 2, 4, and 8 mg kg^{-1} or not:

```
        x      y   Cd Below2 Below4 Below8
1  181072 333611 11.7  FALSE  FALSE  FALSE
2  181025 333558  8.6  FALSE  FALSE  FALSE
3  181165 333537  6.5  FALSE  FALSE   TRUE
4  181298 333484  2.6  FALSE   TRUE   TRUE
5  181307 333330  2.8  FALSE   TRUE   TRUE
6  181390 333260  3.0  FALSE   TRUE   TRUE
7  181165 333370  3.2  FALSE   TRUE   TRUE
8  181027 333363  2.8  FALSE   TRUE   TRUE
9  181060 333231  2.4  FALSE   TRUE   TRUE
10 181232 333168  1.6   TRUE   TRUE   TRUE
11 181191 333115  1.4   TRUE   TRUE   TRUE
12 181032 333031  1.8   TRUE   TRUE   TRUE
13 180874 333339 11.2  FALSE  FALSE  FALSE
14 180969 333252  2.5  FALSE   TRUE   TRUE
15 181011 333161  2.0  FALSE   TRUE   TRUE
16 180830 333246  9.5  FALSE  FALSE  FALSE
```

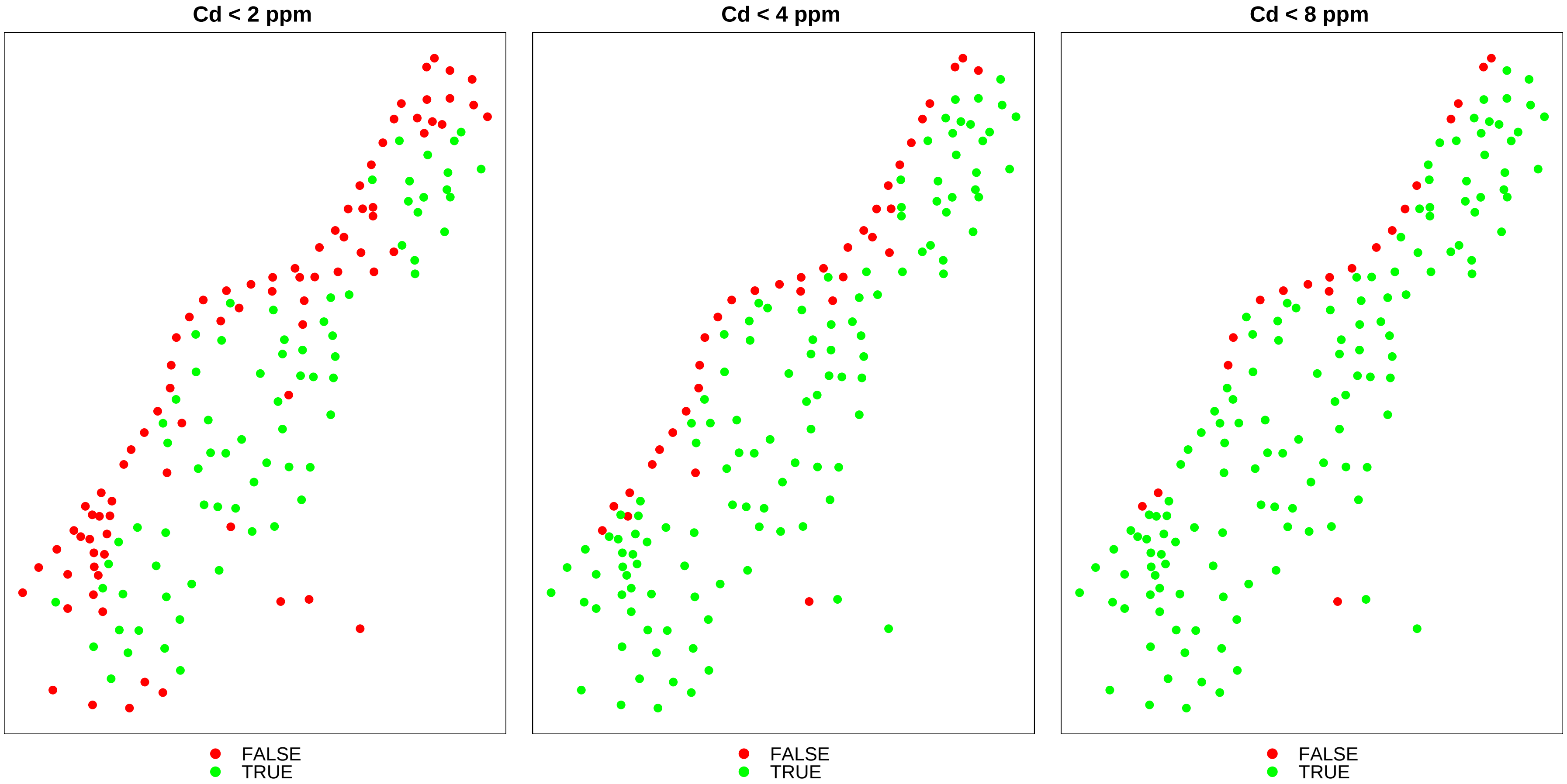
To check your understanding . . .

Q1 : *What happens to the number of* `TRUE` *indicators as the threshold increses?* *Jump to A1* •

Indicator map

- Every sample point is either 1 (‘in’) or 0 (‘out’); a binary map
- No measure of ‘how far’ in or out
- Prepare a **series of indicator maps**, with increasing thresholds, to visualise the **cumulative sample distribution**
- A common strategy is to divide the range of the sample values into **quartiles** or **deciles** and prepare an indicator for each
- The proportion of 1’s will **increase** with increasing quantile.

Indicator maps for three cutoffs



Topic 3: Indicator variograms

- Compute as for a parametric variogram; every sample point has either value 1 (below the cutoff, in the set) or 0.

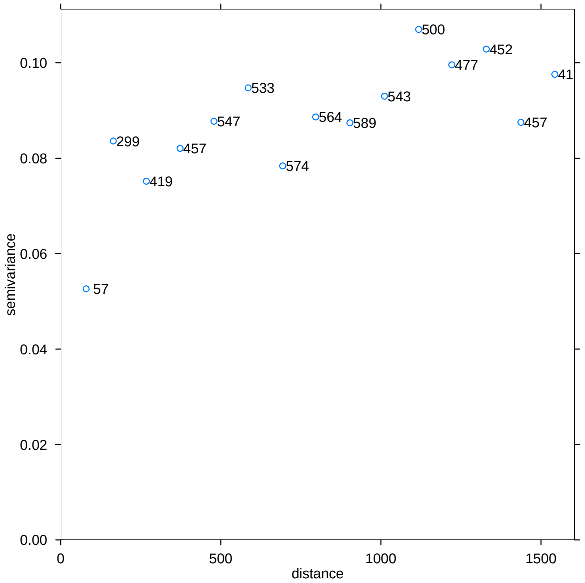
- The semivariance of each point pair is either 0 (both above or below; both out or in) or 0.5 (one above, one below; one out, one in).

- For a quantized continuous variable, each indicator variable (quantile) might have **different spatial structure**

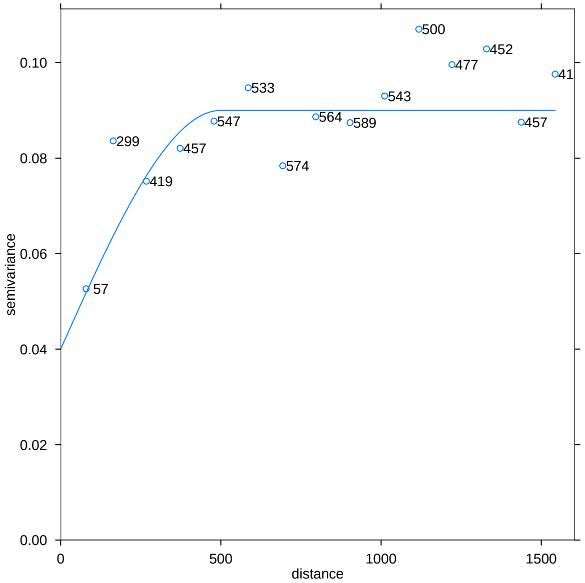
- Variograms near the two ends of the CDF have few 1's or 0's (depending on the end), so few point-pairs will have semivariance 0.5 → hard to model (fluctuates)

- Model as for parametric variogram; however **by theory** the total sill must be < 0.25 (generally it's a lot lower, except near the median)

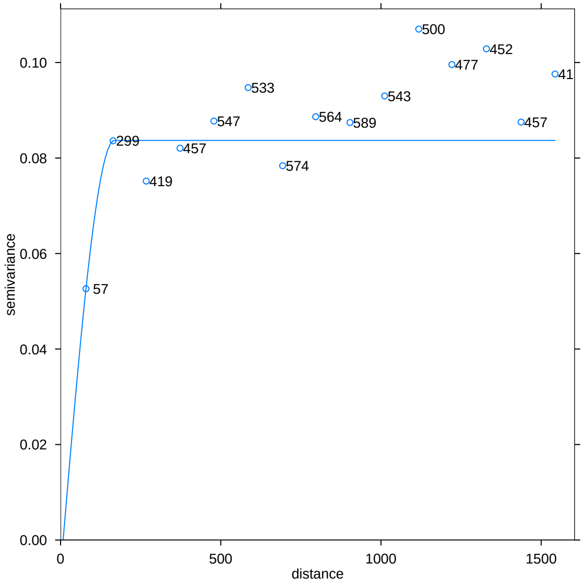
Indicator variogram (9th decile);



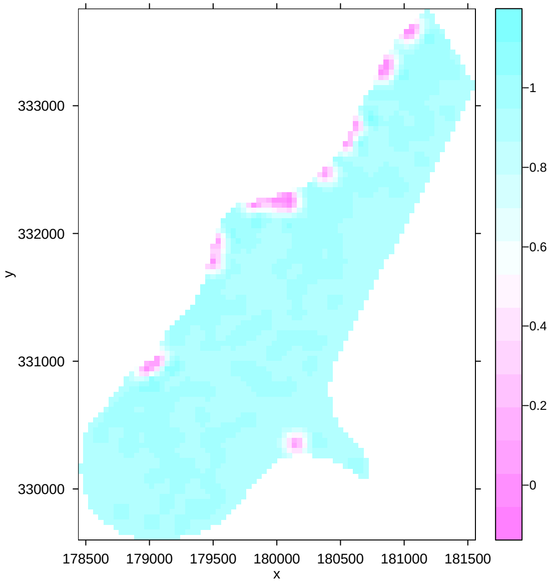
Estimated model



Fitted model
note unrealistic nugget



Probability < 8.26mg kg^{-1}



Topic 4: Kriging with indicator variables

This is a simple **non-parametric** (also called **distribution-free**) method of **prediction**.

Three types of maps are possible, depending on objective:

1. **Probability** that each point is below a defined threshold: **probability kriging**
   - This can be used directly in hazard mapping, if the threshold is set to represent the hazard level.
2. An entire **cumulative probability distribution** (CDF) at each point.
3. Predicted **value** at each point (as in OK or UK).

The first is useful in e.g. pollution studies. The other two are non-parametric alternatives to parametric OK in the presence of outliers. We will only look at the first.

Probability kriging using indicator variables

1. Calculate the **indicator variable** at the required **threshold**

2. Calculate the **empirical variogram** for that indicator

   - Note: May have to use a threshold closer to the median if there are too few 1's so that the variogram is erratic.

3. **Model the variogram**; note total sill should be < 0.25.

4. **Solve the kriging system** at each point to be predicted using OK or SK (see next slide)

   - If necessary, **limit the results** to the range $[0 \ldots 1]$

5. Interepret this map as the **probability** that the point does **not** exceed the selected threshold.

Estimating the mean probability with IK

There are three ways to estimate the mean probability:

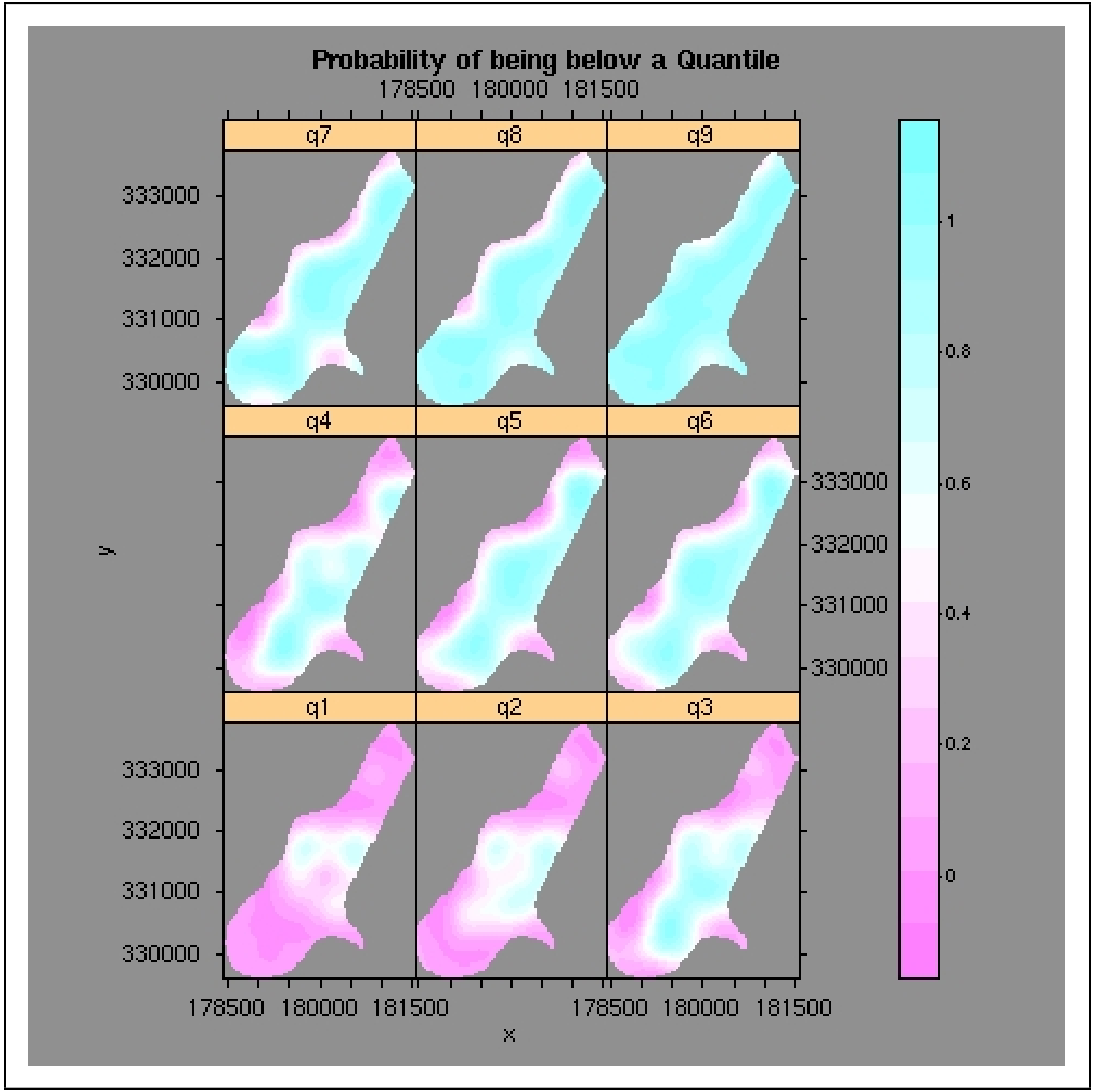
1. From the sample set itself **during** the kriging process, i.e. using OK or a variant such as UK;

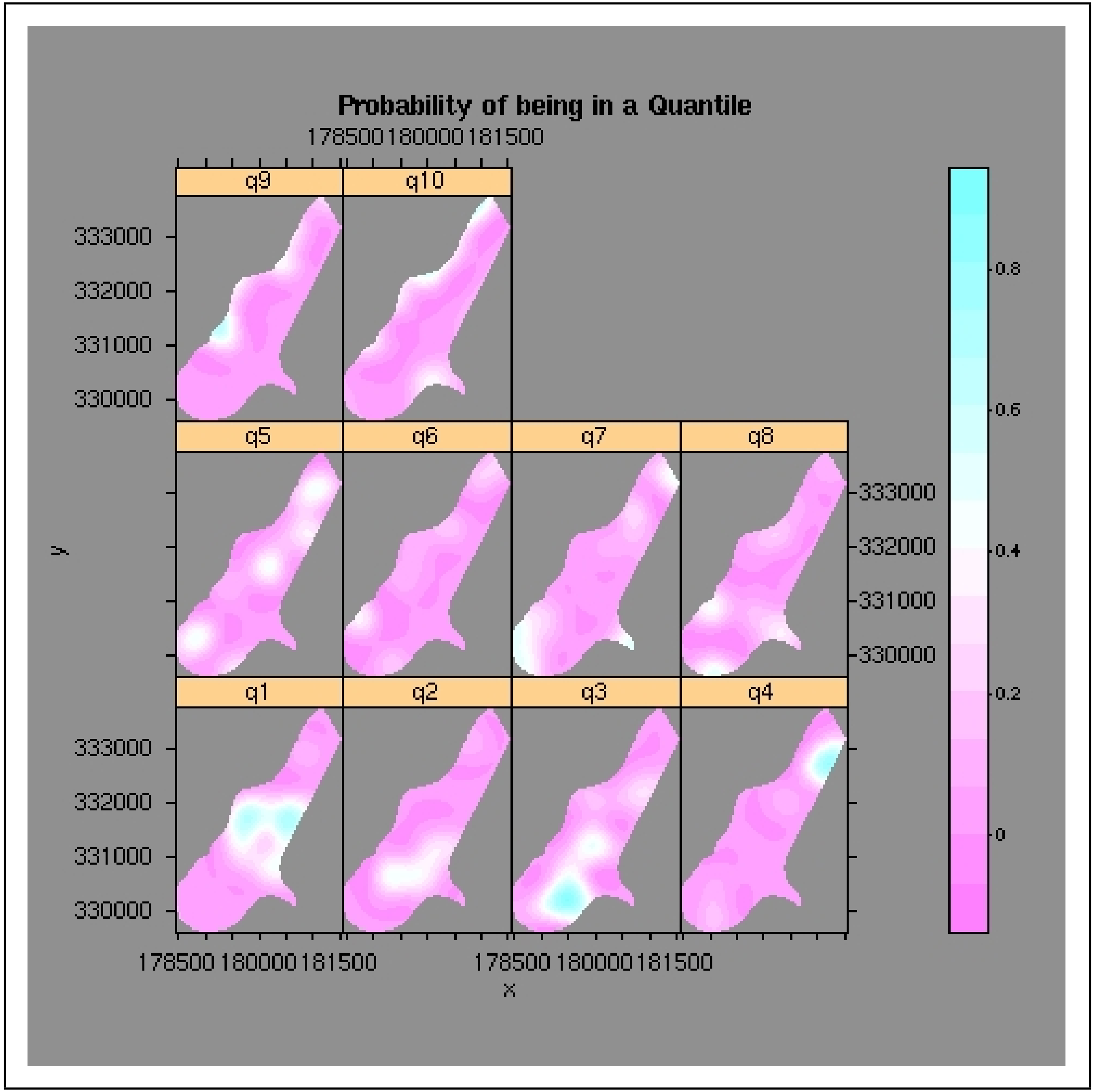
2. From the sample set itself **before** the kriging process, from the **indicator proportion** in the sample set;

3. From *a priori* knowledge (e.g. previous studies); this is a good idea if the sample is biased in some way, e.g. towards suspected polluted sites.

The second method uses **Simple Kriging (SK)** with the **indicator proportion as the expected value**.

- Recall: SK does not estimate the mean, instead it is supplied by the analyst.

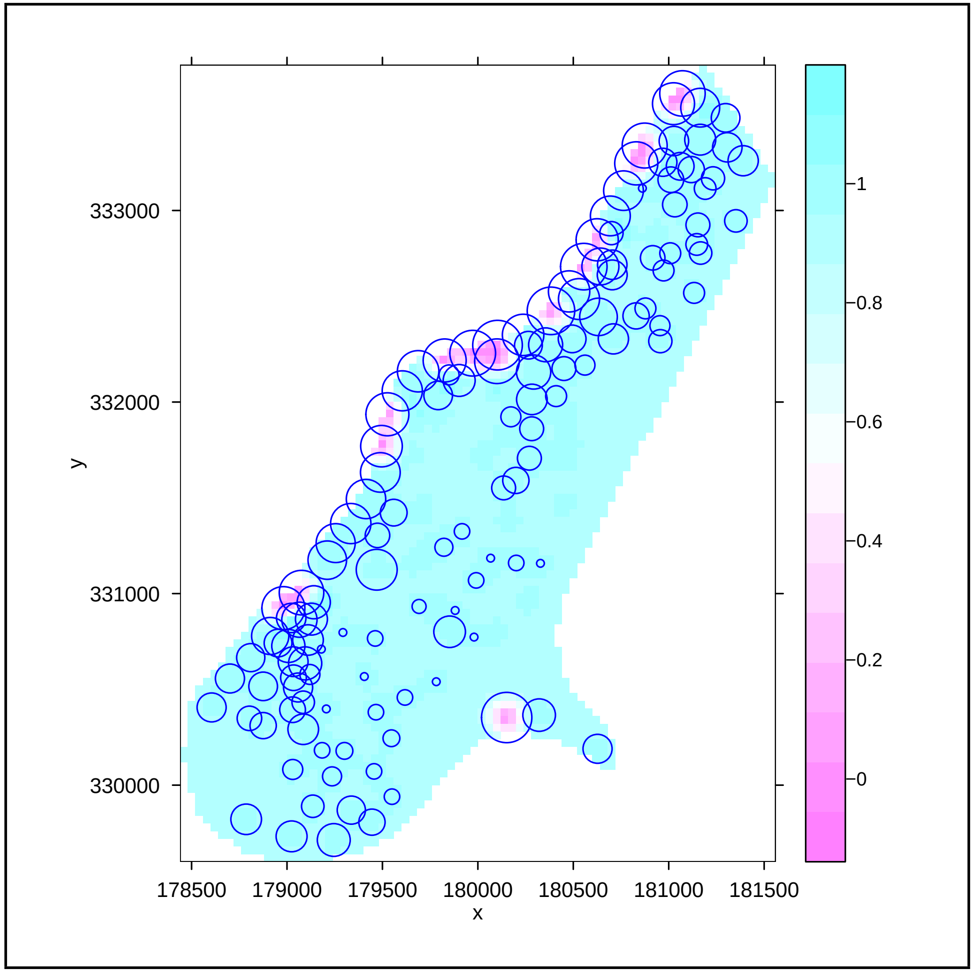

Probability of being below a Quantile
q7
q8
q9
q4
q5
q6
q1
q2
q3
178500
180000
181500
333000
332000
331000
330000
y
x
1
0.8
0.6
0.4
0.2
0

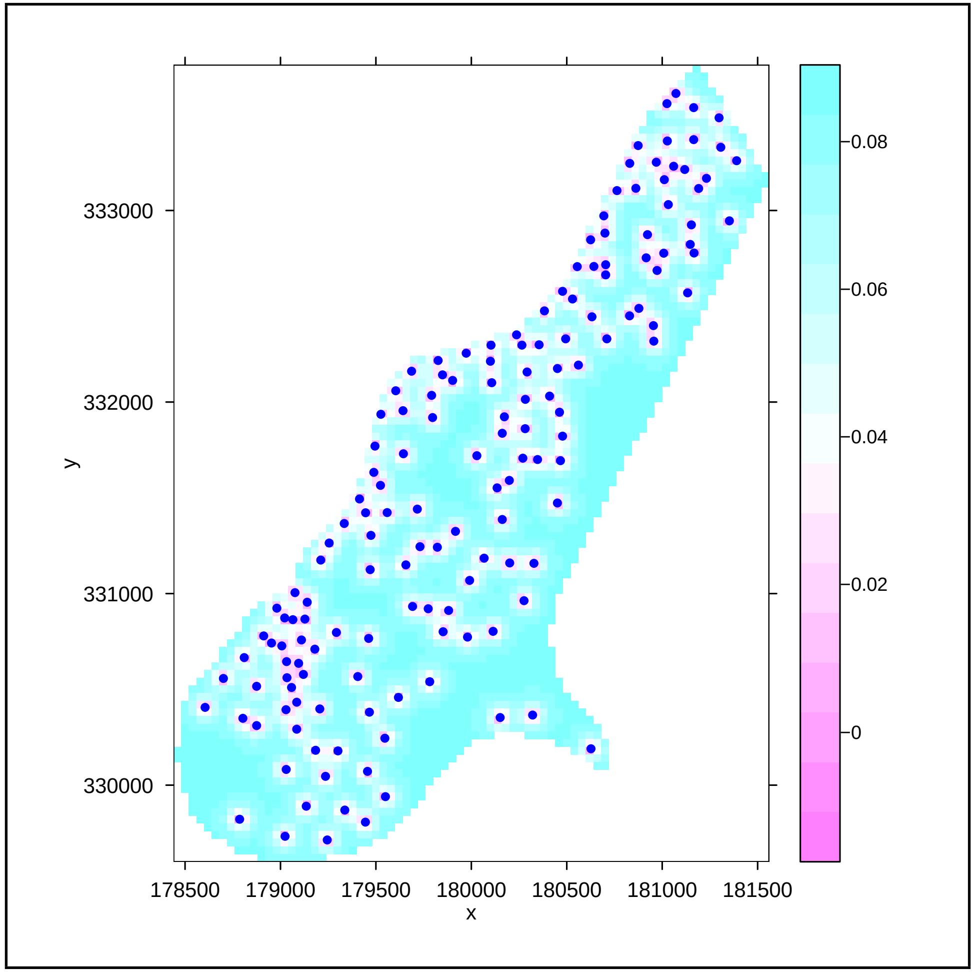

Probability of being in a Quantile
178500 180000 181500
q9
q10
q5
q6
q7
q8
q1
q2
q3
q4
333000
332000
331000
330000
y
x
178500 180000 181500
0.8
0.6
0.4
0.2
0

Intepretation of the kriging variance for probability kriging

- The IK predicted ‘value’ is already a probability!

- So the IK kriging variance is the variance of a probability . . . whatever that means





Probability (with data points)

Error variance of this probability

Summary: Advantages of IK

- Makes **no assumption about the theoretical distribution** of the data values, yet still give **realistic probability estimates**

- **Outlier-resistent**: these can not increase the estimate or kriging variances of an indicator arbitrarily; for data values they only affect one quantile

- Simple Kriging may be used at each quantile, which improves the estimate.

Summary: Disadvantages of IK

- Unsound theoretical basis in many cases;
- **Variograms may be difficult to model**, especially at the highest and lowest quantiles (few pairs with different 0/1 values);
- For estimating values: this combines probabilities computed with different variograms for different quantiles of the same variable; so a single median variogram is used, but is this correct for each quantile?;
- Problem of the meaning of indicator predication variance maps: a probability of a probability means . . . what?

Further reading

Isaaks, E. H., and R. M. Srivastava (1990), An introduction to applied geostatistics, Oxford University Press, New York. Chapter 18: "Estimating a distribution"

Goovaerts, P., and A. G. Journel (1995), Integrating soil map information in modelling the spatial variation of continuous soil properties, European Journal of Soil Science, 46(3), 397-414.

Goovaerts, P., R. Webster, and J. P. Dubois (1997), Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics, Environmental and Ecological Statistics, 4(1), 31-48.

Lark, R. M., and R. B. Ferguson (2004), Mapping risk of soil nutrient deficiency or excess by disjunctive and indicator kriging, Geoderma, 118(1-2), 39-53.

Papritz, A. (2009), Why indicator kriging should be abandoned, Pedometron, 26, 4-7.

Answers

Q1 : *What happens to the number of `TRUE` indicators as the threshold increses?* •

A1 : *They increase; more of the CDF is below the specified value.* *Return to Q1* •