PSCS/NTRES 6200
Spatial Modelling and Analysis

Concepts of Space and Spatial Modelling

D G Rossiter

Cornell University
Section of Soil & Crop Sciences

January 21, 2018

Topic: Types of "spaces"

- The word **space** is used in mathematics to refer to any set of variables that form metric axes and which therefore allow us to compute a **distance** between points in that space;
  - If these variables represent **attributes**, we have a **feature space**.
  - If they represent geographic **coördinates**, we have a **metric geographic** space;

- A non-metric mathematical space can represent topology; if these are geospatial **relations**, we have a **topological geographic** space.

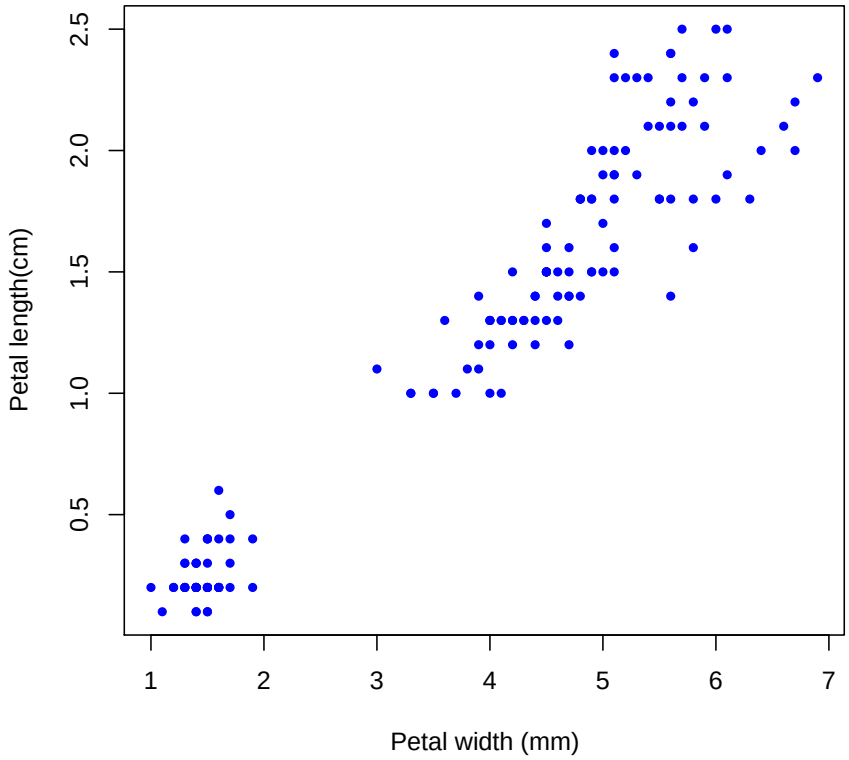
Feature space

This “space” is the metric space (in the mathematical sense) formed by any set of variables:

- **Axes** are the range of each variable;
- **Coördinates** are values of variables, possibly transformed or combined;
- The observations are related in this ‘space’, e.g. the “distance” between them can be calculated.
- We often plot variables in this space, e.g. **scatterplots** in 2D or 3D.
- This is the basis of bi-, tri-, multi-variate analysis

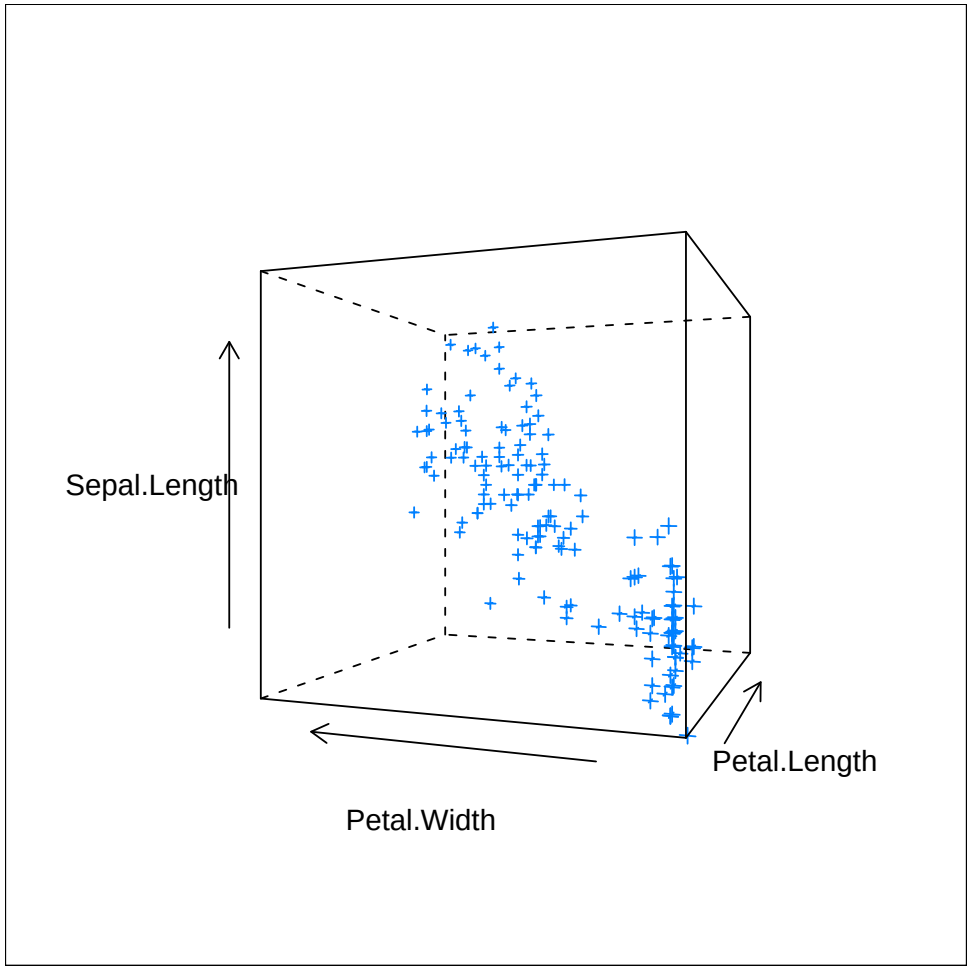
Note: **Feature space** is sometimes referred to as **attribute space**.

Scatterplot of a 2D feature space



This is a **visualisation** of a 2D feature space using a **scatterplot**; two attributes of individual iris flowers as coördinates.
Source: Anderson, Edgar (1935), *The irises of the Gaspé Peninsula*, Bulletin of the American. Iris Society, 59, 2-5.

Scatterplot of a 3D feature space



Anderson *Iris* data, three attributes as coördinates.

Geographic space

- "Geo" + "graphy" = "Earth" + "mapping"
- Related somehow to the Earth's surface
- **metric** vs. **topological**

Metric geographic space

- a mathematical space where the axes are **map coördinates** that relate points to some reference location on or in the Earth (or another physical body)

- These coördinates are often in some **geographic coördinate system** that was designed to give each location on (part of) the Earth a unique identification; a common example is the Universal Transmercator (UTM) grid.

- However, a **local coördinate system** can be used, as long as there is a clear relation between locations and coördinates;
  - ungeoreferenced aerial or satellite imagery
  - photograph of microscope slide (not the Earth's surface but has metric geometry)

- Key point: can compute **distances**, **angles** of separation, and **areas**.

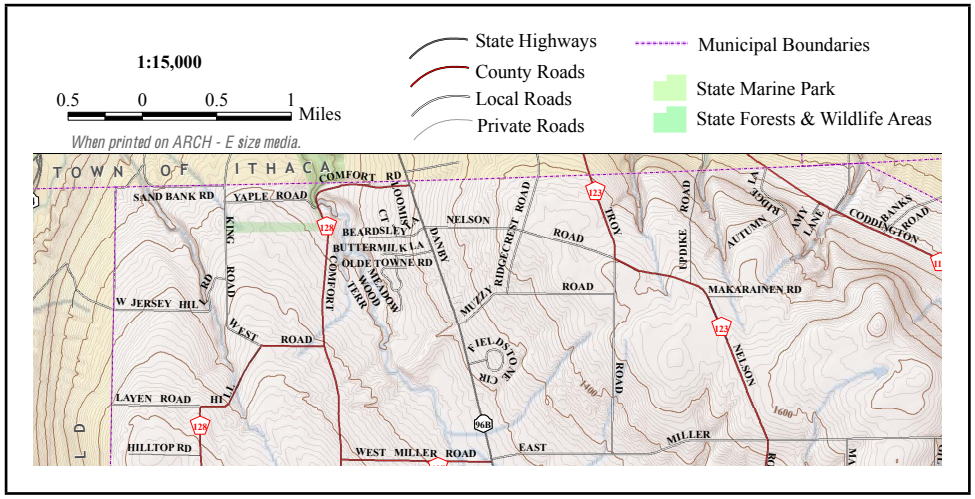
Metric space

- Coördinates represent "true" distances along their axes
- Axes are 1D **lines**; they almost always have the **same units of measure** (e.g. metres, kilometres . . .)
- **One-dimensional**: coördinates are on a line with respect to some origin (0): $(x_1) = x$
- **Two-dimensional**: coördinates are on a grid with respect to some origin $(0, 0)$: $(x_1, x_2) = (x, y) = (E, N)$
  - **Latitude-longitude** (sometimes called "geographic") coördinates do not have equal distances in the two dimensions; they should be transformed to metric (grid) coördinates for geo-statistical analysis.
- **Three-dimensional**: coördinates are grid and elevation (or depth! a negative elevation) from a reference elevation: $(x_1, x_2, x_3) = (x, y, z) = (E, N, H)$

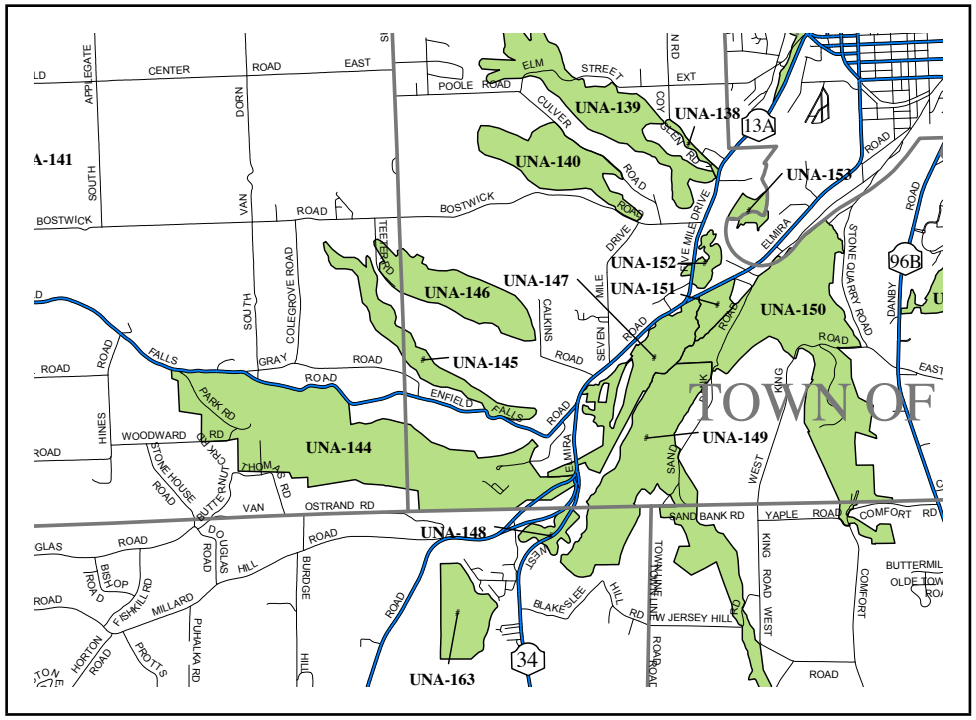
Maps of metric space

- Shows features as **abstract** objects (points, lines, polygons, grids);
- The features are **labelled**; these are collected in a **legend**;
- A map shows both **metric geographic** and **feature** (attribute) spaces;
- One map can show many **attributes**
  - Special case: contour maps “2.5D”, show the third geographic dimension (height, depth) on a 2D display

Some maps in metric space

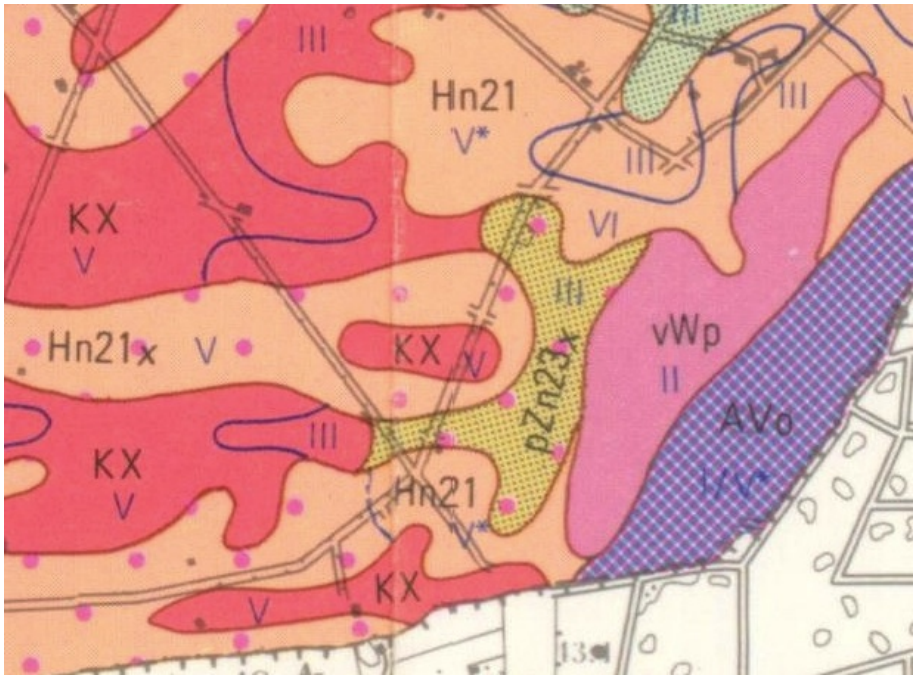


contours; themes: roads (lines); parks & administrative divisions (polygons)

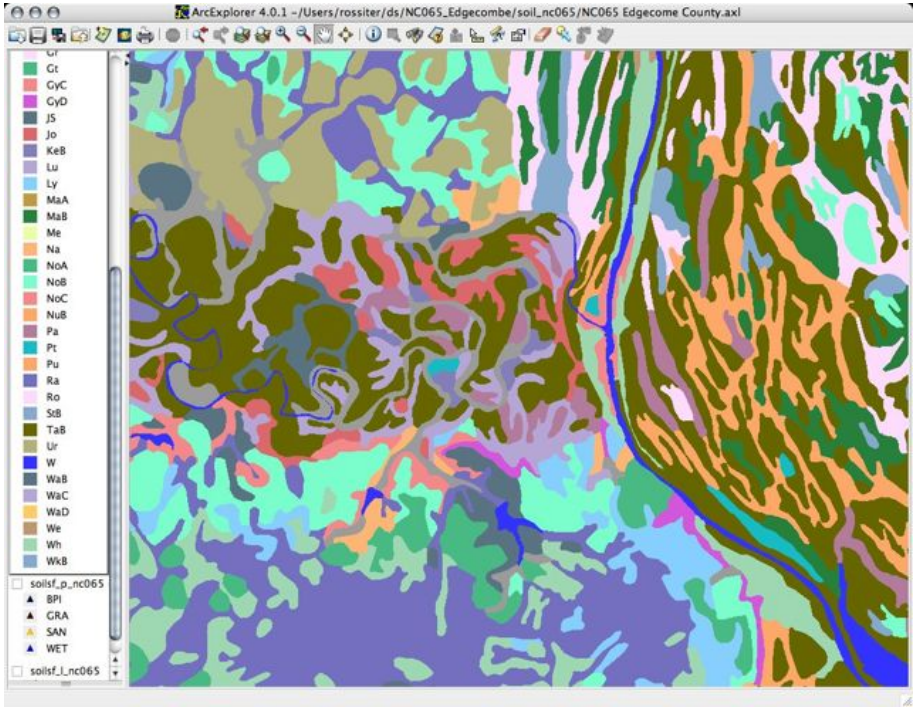


labelled polygons and lines

Soil class maps

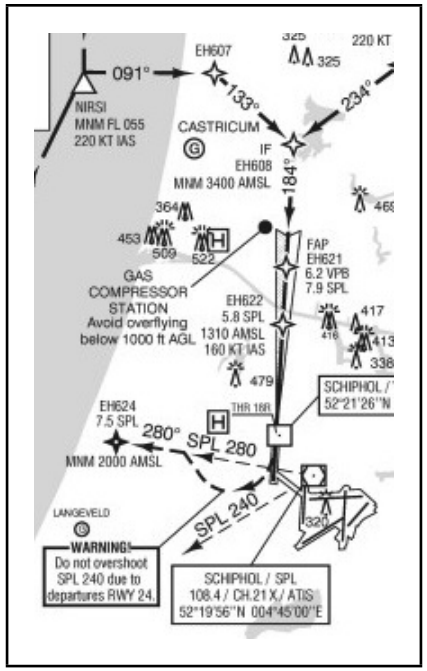


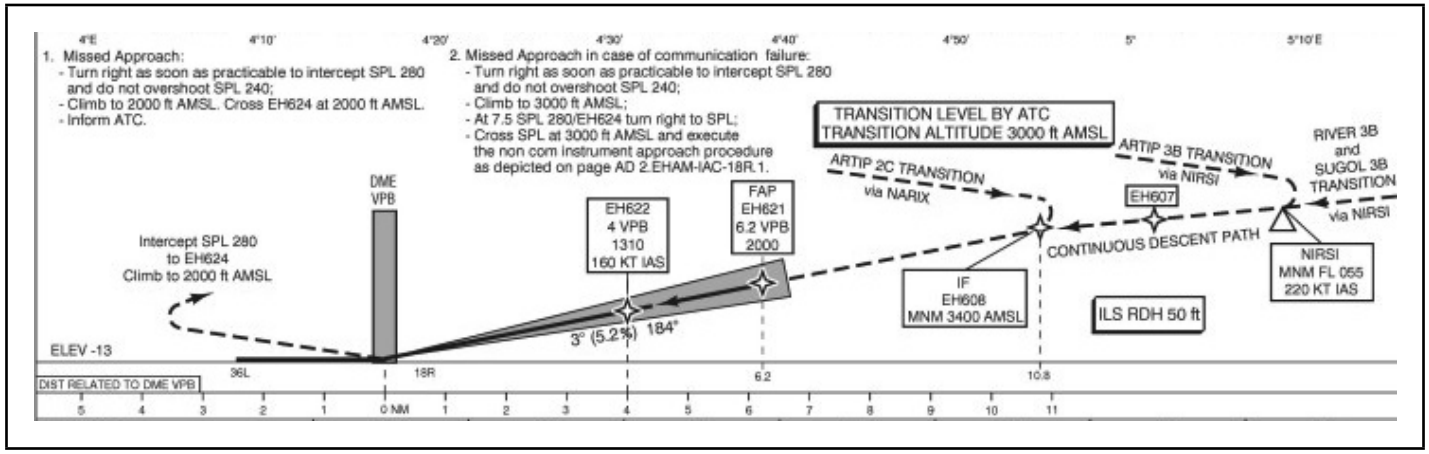
Dutch soil survey 1:50k sheet 34E; themes: soil class; ground-water level class; subsoil condition ("polkadot" overprint)



SSURGO 1:25k, Edgecombe County NC; theme: soil mapping units
n.b. extremely poor, non-connotative, colour scheme

Navigation maps – 1D and 2D

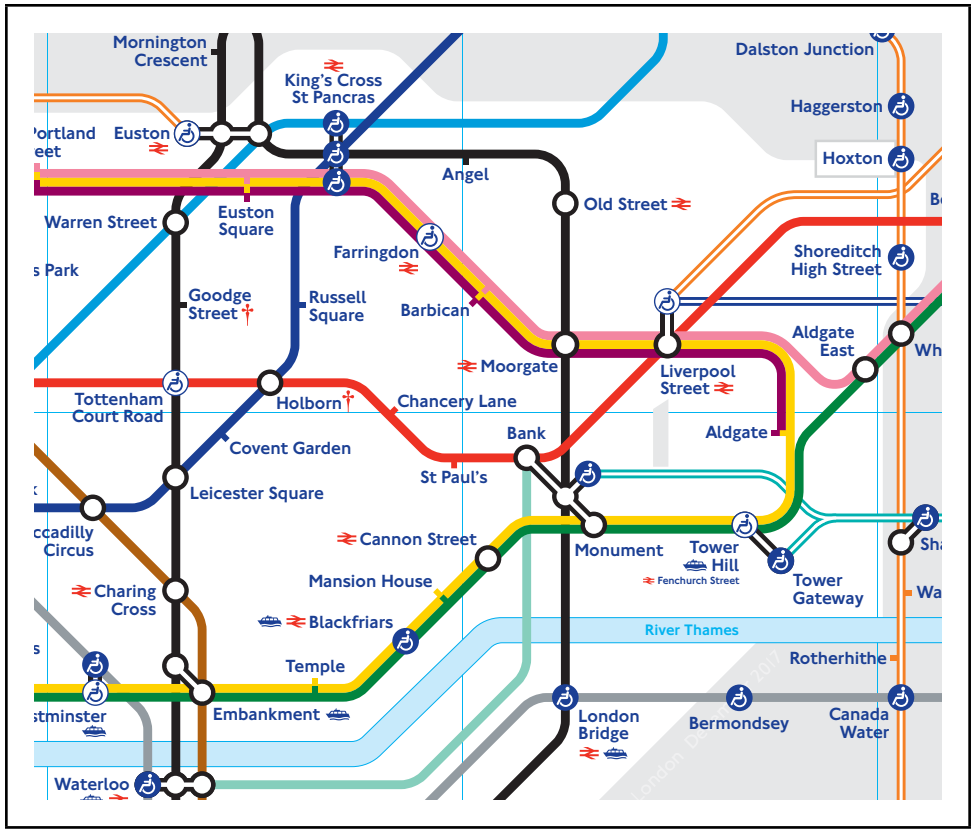


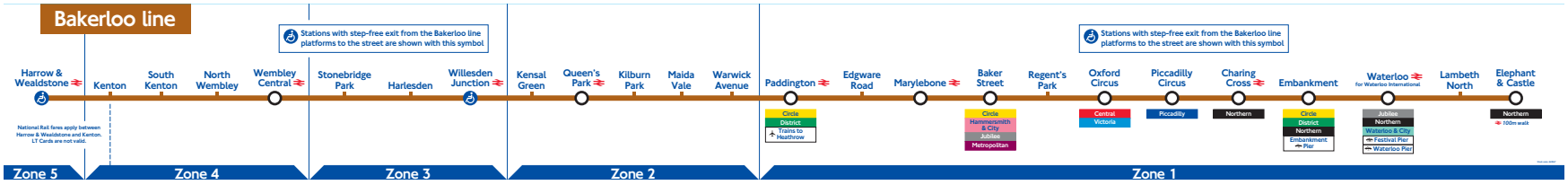


Topologic geographic space

- Topologic relations are preserved;
  - adjacency, connectivity, containment, intersection . . .
- Distances are not true

True 1D, 2D topology, distorted distances and angles





Note: 3D not correct topology

Topic: Thinking about spatial analysis

Reference: O'Sullivan, D., & D. Unwin. 2010. *Geographic information analysis*. 2nd ed. Wiley

Cornell access to e-book:

`http://resolver.library.cornell.edu/cgi-bin/EBookresolver?set=Books24x7&id=35218`

Their title term, admittedly "a rather new concept", is, I think, covered by the common meaning of the existing term **spatial analysis**, which they use in another sense (see following).

O'Sullivan & Unwin's classification

Four concepts:

1. Spatial data **manipulation**

2. Spatial data **description and exploration**[1]

3. Spatial **statistical analysis**

   - Can a statistical model represent the data?
   - This is not yet understanding, only summarizing as an empirical relation.
   - Requires special techniques to account for spatial relations

4. Spatial **modelling**

   - Understand **functional form** of spatial **processes**
   - **Predict** spatial outcomes

[1] O'S & U call this "spatial data analysis"

Typology of spatial data – views of the world

(O'Sullivan & Unwin – modified)

Objects real-world **entities**: can be **discretely** identified "in the field" and located in geographic space

Fields **continuously-varying** properties in space

- Represented in a GIS by some **discretization**, but conceptually continuous
- Measured with some **spatial support** (sample size, instrument field of view, . . .), but conceptually continuous

Networks **interconnected** line and point objects

It's not so simple ...

- the conceptual definition of the object may be vague (fuzzy **definition**)
- It may be difficult to identify an object in the field (fuzzy **identification**)
- Objects may have **fuzzy boundaries or locations**
- Object concepts may depend on the map **scale**
  - Roads are conceived of as polygons at large map scale, lines at small map scale
  - Buildings are conceived of as polygons at large map scale, lines at small map scale
- Continuous variables are measured with some **spatial support** (sample size, instrument field of view, ...); this is a lower limit of the **resolution** with which we can describe the conceptually-continuous field

Typology of spatial data – conceptual models of spatial objects

Point a single point location, defined by coordinates

Line a set of ordered points, connected by straight line segments

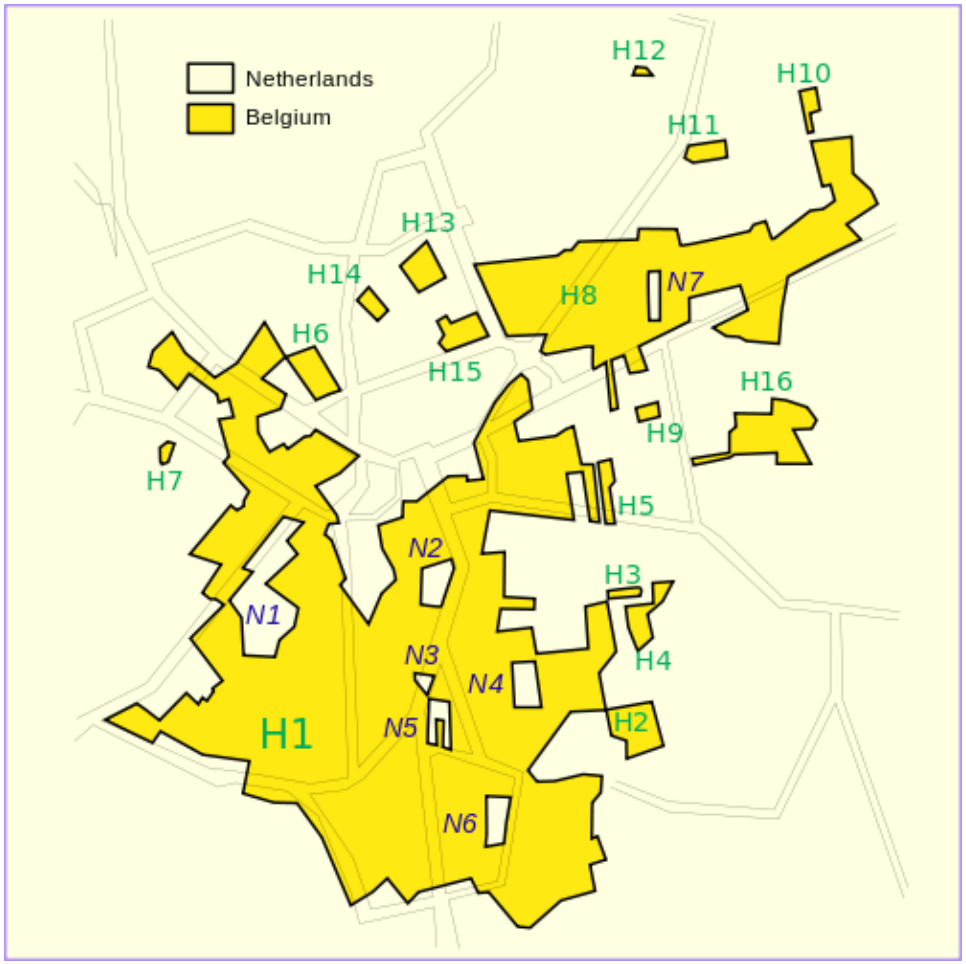
Curves a set of control points (not necessarily on the curve); and a mathematical function of coordinates (e.g., splines)

Polygon an area delineated by one or more lines, possibly containing holes (and holes within holes . . .)

Network a group of curves connected at points (**vertices**); mathematical "graph"

Grid a collection of points or cells, organised in a regular lattice covering an area

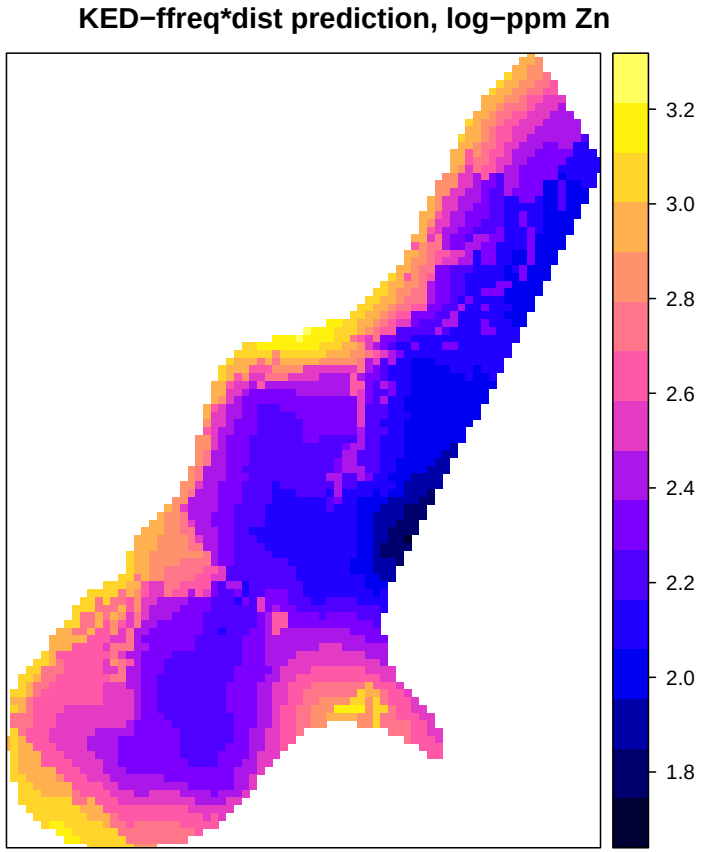
Polygons

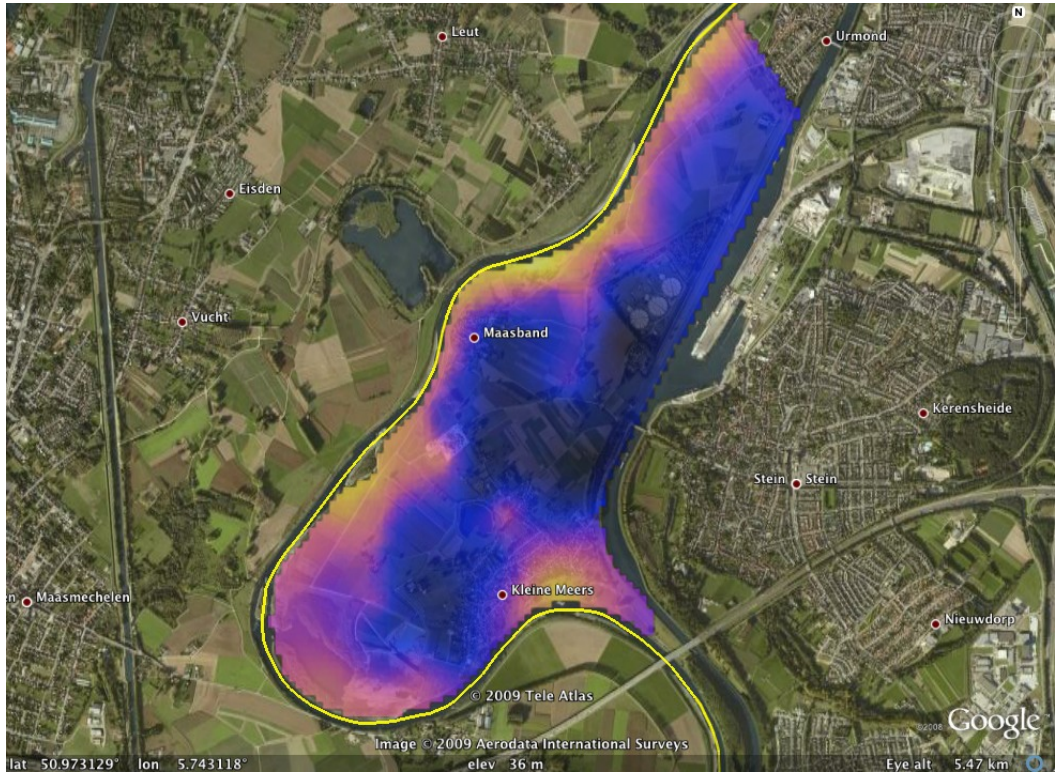


Baarle-**N**assau (NL), Baarle-**H**ertog (*Baerle-Duc*) (B); source Wikimedia commons

"H" polygons are B inside NL (i.e., holes in the surround unlabelled "N"); "N" polygons are N inside B (i.e., holes in a "H" polygon).

Grids





40 x 40 m square cells

conceptually smooth

Cell values could be centre point predictions, block averages, block maxima . . .

Typology of spatial data – data models

These are how spatial data are represented **inside a GIS** – *not* their conceptual representation

Vector exact mathematical form: 0-dimension = points; 1-dimension = line segments (which can be joined); 2-dimension = areas; 3-dimension = volumes

- Note: a Triangulated Irregular Network (TIN) is a vector data model of a 2-D continuous surface conceptual model

Raster a regular tessellation (e.g., square or hexagonal grid); fixed resolution

- data values may be grid cell averages, maxima, minima . . . or single values at the centre point

Topic: Spatial analysis & modelling

- Abstracting and modelling some aspect of a **spatial reality**
  - Natural resources
  - Built environment
  - Social environment
  - Conceptual enviroment (e.g., political divisions)

- Does **not** include modelling objects in space without somehow considering their spatial position, i.e., pure feature-space analysis
  - Just displaying the results of a feature-space model on a map does not make a spatial analysis

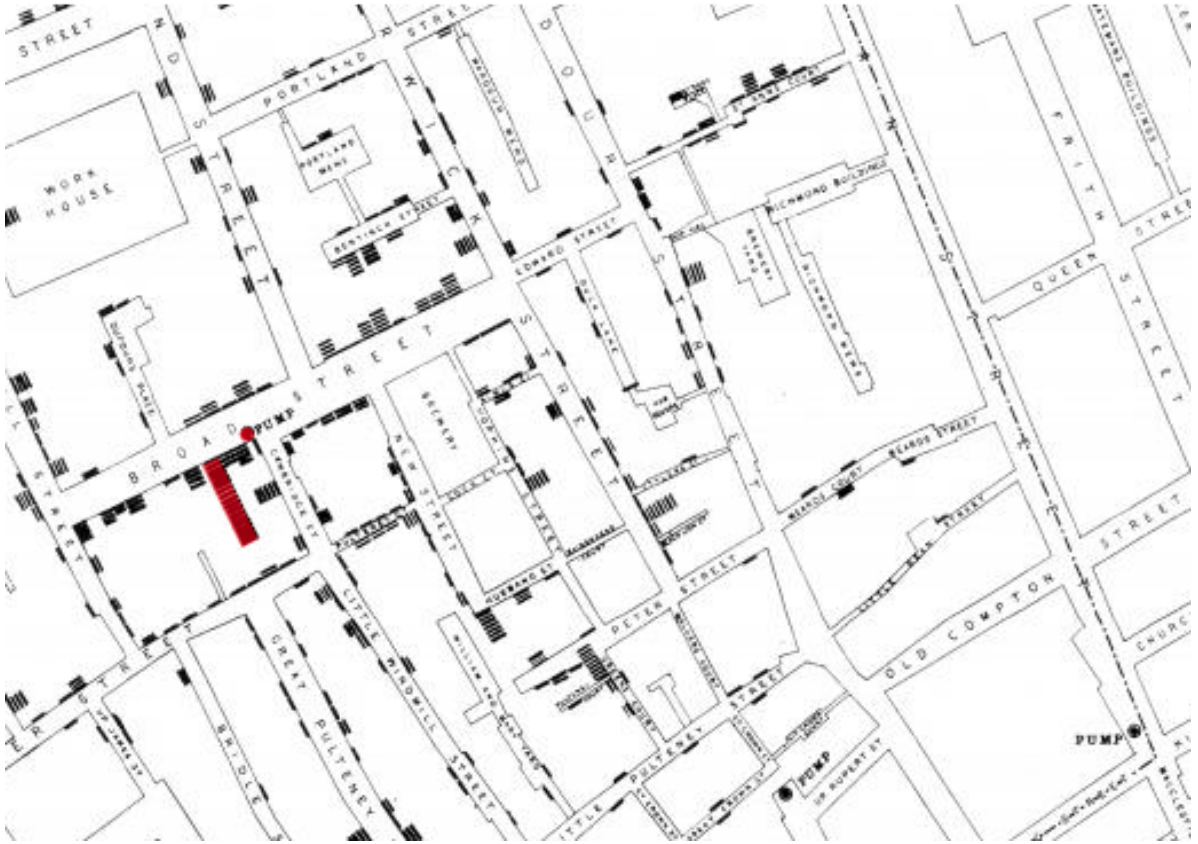
Why model?

- **hypothesis formulation**: a successful model suggests that it realistically represents some real process
- **hypothesis testing**: if we can reproduce some spatial phenomenon with our model based on the hypothesis, the support for the hypothesis is increased
  - any complicated hypothesis is not tested as such, we build up evidence to support, refute or modify it
- **understanding** of "nature": a successful model increases our confidence that the model structure matches the real structure
- spatial(-temporal) **prediction**: the model results in a map which is then used for **decision-making**
- **scenario** analysis: "what if. . . ", mainly for decision-making under **uncertainty**

Example – hypothesis formulation and testing

Observation: clustering of cholera, relation to water sources

Hypothesis: cholera is an infectious disease, caused by an organism which lives in wastewater and cycles through humans.



source: Snow, John. *On the Mode of Communication of Cholera*, 2nd Ed, John Churchill, London, 1855

Types of models – 1

Physical capture the essential behaviour of a physical system with equations; also called **mechanistic**

Empirical determine relation between system components, without necessarily knowing the cause

In practice the line is blurry:

- Physical principles are often used to motivate choice of variables in empirical models
  - But, pure **data mining** models, e.g., artificial neural networks (ANN), are purely empirical.
- "Physical" models have many parameters that must be empirically calibrated (very rarely deriveable from first principles)

Types of models – 2

Explanatory the main purpose of building the model is to **understand the process** which gave rise to the object of study

- e.g., ecological factors controlling species distribution, based on observations of the species and co-variables
- are not necessarily useful for prediction, but often are

Predictive the main purpose of building the model is to **predict** at unsampled locations / times (especially the future)

- e.g., areas suitable for a proposed land use, basing the model on areas currently more-or-less successful for that use
- do not necessarily lead to understanding, but often do

For pure prediction, a **black-box** model may be acceptable and even perform well (e.g., ANN), but such a model is useless for understanding.

Modelled geographical objects

All of these can be **model inputs** or **model outputs**:

- points (locations, possibly with attributes)
- lines (same)
- polygons / groups of polygons (same)
- continuous fields (described mathematically or discretized)

Feature-space **attributes** linked to the geographic features may be part of the model

Examples of spatial models

- Distribution of rare species in a forest (**point** pattern, no attributes)

  1. relation to spatially-distributed ecological factors (feature space, but distributed in geographic space)
  2. purely spatial relations, e.g. seed dispersal, allelopathy . . .

- Spread of a disease epidemic

  - point cases, possibly with attributes e.g. age, gender, previous health . . .
  - point or line water sources, possibly with attributes, e.g., water quality
  - point pollution sources, possibly with attributes
  - continuous fields, e.g. soil permeability or hydraulic conductivity

- Distribution of pollutants in soil or groundwater (continuous **field**)

- "Optimal" location of a new school (etc.), considering spatial factors

Types of processes being modelled

The idea is to **match** the **model** with a **true process** that caused the observations.
Types of processes:

- **fluxes** (flows) driven by "physics": e.g., diffusion, convection / advection, radiation
  - These can be in "1"D (e.g., along a river, through a road network), 2D (e.g., hillslope), 3D (e.g., soil volume above groundwater)
  - Non-physical processes, e.g., population migrations, could be modelled by physical equations, if assumptions are met
- **known processes**: e.g., plant growth affected by heat, light, nutrients, competition . . .
- **population dynamics**: e.g., birth / death, preditor / prey, cooperation / competition

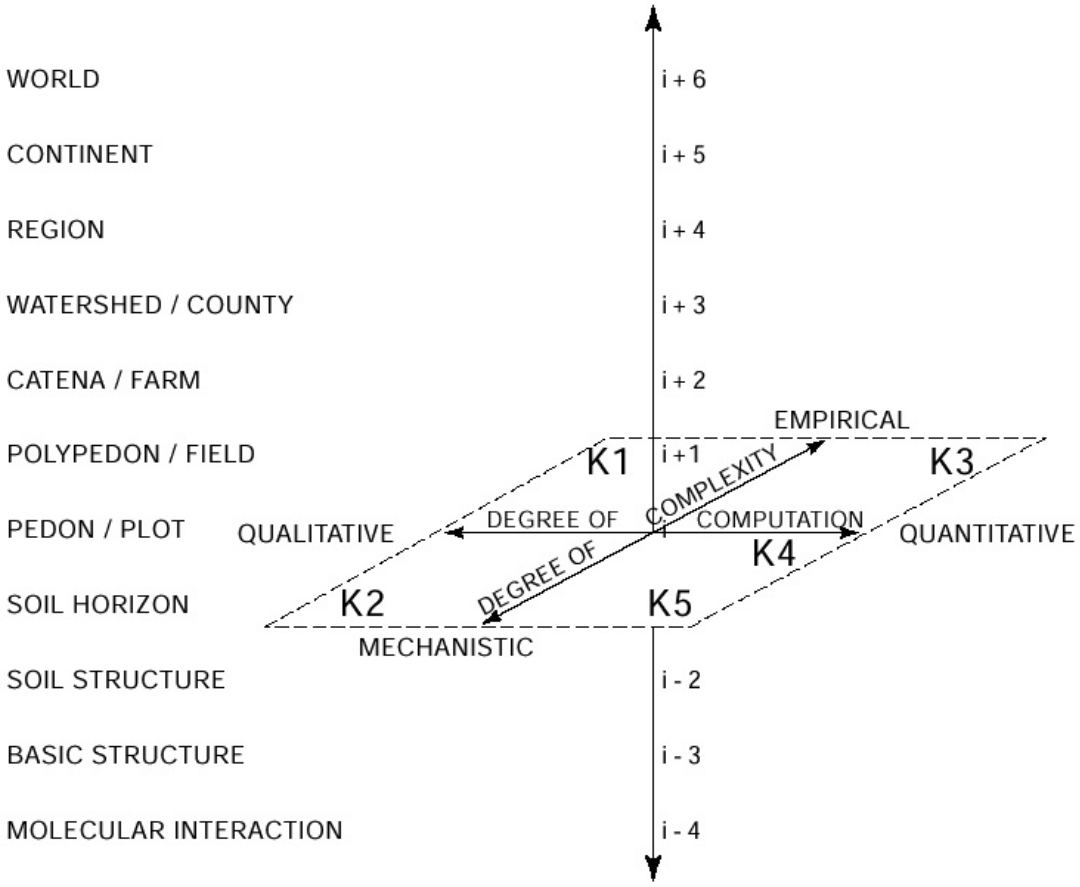
(continued . . .)

Processes (continued)

- **"intelligent" agents** making decisions and interacting

- **decisions** (maybe under uncertainty): try to reproduce the decision-maker's logic and criteria (e.g., site selection)

Three-axis model classification



after Bouma (1999) *Land evaluation for landscape units* In Sumner M. E. (Ed.), Handbook of soil science (pp. E393- E412). Boca Raton, FL: CRC Press

Based on Hoosbeek, M. R., & Bryant, R. B. (1992). Towards the quantitative modeling of pedogenesis – a review. Geoderma, 55, 183-210.

Most models have components spread through this diagram. They must be linked, but this brings many problems of concepts / scales.

Three-axis classification

- **degree of complexity**: Mechanistic vs. empirical – see above
- **scale** – see below
- **degree of computation**: Qualitative vs. quantitative

Degree of computation: algorithm / outputs more or less **quantified**

- e.g., "highly suitable" vs. "Net present value for intensive vegetable production \$1000 ha^{-1}"

Concepts of "scale"

1. **cartographic** (map) scale: relation of map distances to ground distances
   - "large" = large area of paper needed to represent a given ground area
2. **geographic** scale: size of area being studied
   - "large"= over a wide area
3. **process** scale: spatial extent / variability of process operating on landscape
   - e.g., soil erosion: rill, plot, small catchment, river system . . .
4. **measurement** (observation) scale: size ("support") of observations
   - e.g., soil ped, core, profile, pit, trench, . . .
5. **modelling** scale: size of fundamental area at which processes or objects are represented in models ("support")

Blöschel, G., & M. Sivapalan. 1995. *Scale issues in hydrological modelling: a review*. Hydrol. Proc. 9: 251-290.

Temporal scale

The above-mentioned **spatial** scales can also be used to describe **temporal** scales.

(Except for “cartographic”).

Matching scales

Key points:

- **modelling** scale should match **process scale**
- **information** at different measurement scales must be **harmonized**
  - e.g., satellite imagery at different resolutions
  - e.g., demographic information at census ward vs. postal code vs. administrative unit (these also at different levels)
  - up-, down-scaling (see below)

Up- and down-scaling

Upscaling from detailed scale (e.g., lab. experiments) to coarse scale (e.g., ag. field, region, continent . . .)

Downscaling from a coarse scale (e.g., general circulation model of the atmosphere) to a detailed scale (e.g., local weather forecast)

These can be either **spatial** or **temporal** scales.

Often the **inputs** to a model do not have the same scale, so some must be adjusted; and/or the input scale does not match the desired **output** scale.

Issues in spatial scaling

- **Up**scaling: must compress/summarize information
  - e.g., area weighted averaging of properties – but is this meaningful? (e.g., white car in black parking lot → grey pixel)
  - maybe re-run models with coarser scale inputs
  - maybe interpolate including information (somewhat) outside the upscaled resolution

- **Down**scaling: must create new **spatially-explicit** information at a finer scale
  - Just increasing pixel resolution is not creating information!
  - Example: disaggregating a coarse-resolution soil polygon (soil **association** with known landscape relation) to fine-resolution polygons (soil **consociation** = more-or-less homogeneous unit); using expert knowledge + covariates (e.g., terrain classification)

- Bui, E. N., & Moran, C. J. (2001). Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma, 103(1-2), 79-94.
- Khan, M. R., de Bie, C. A. J. M., van Keulen, H., Smaling, E. M. A., & Real, R. (2010). Disaggregating and mapping crop statistics using hypertemporal remote sensing. International journal of applied earth observation and geoinformation, 12(1), 36-46.

The modifiable areal unit problem

Issue: the **same analysis** may give **different results**, i.e., lead to different inferences / conclusions, if the data is aggregated at **different scales**

Example: voting patterns by large → small geographic area (polygon size):

- Vote for B.H. Obama vs. W.M. Romney, US President, 2012:

  1. USA (Obama 51.4%)
  2. NY state (Obama 62.6%)
  3. NY 23rd congressional district (Romney wins, Obama 48.4%)
  4. Tompkins County (Obama 68.2%)
  5. City of Ithaca (Obama 83.3%)
  6. 5th ward (Obama 85.2%)
  7. 5th ward 2nd district (Obama 91.6%),

- Summarize predictor variables at the same scales (party registration, census ethnicity, IRS-reported income . . .)

- **Do you expect the same predictive model?**

Spatial modelling

Several attempts have been made to categorize the **conceptual operations** in spatial models, e.g.:

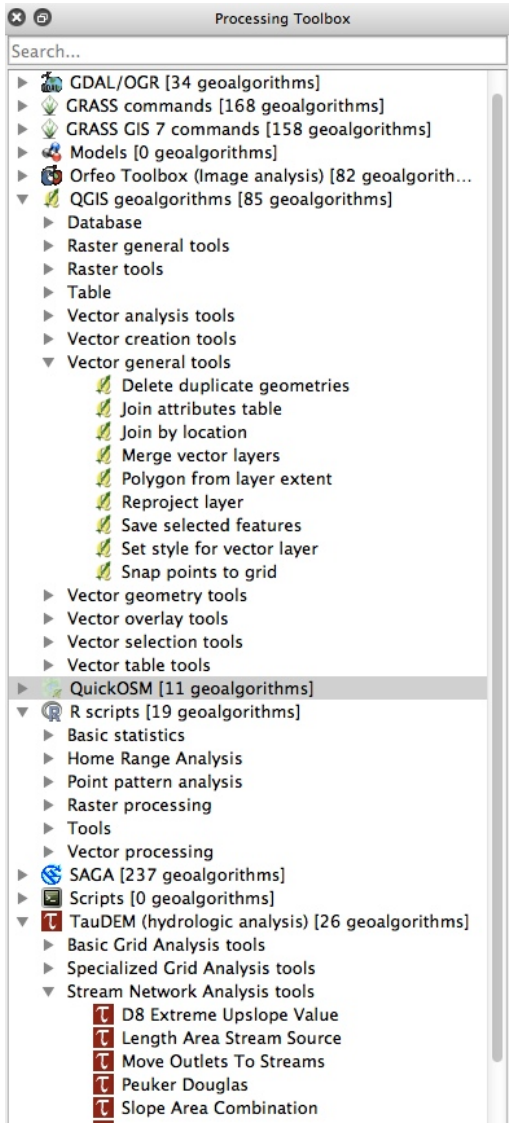
- Burrough, P. A., & McDonnell, R. A. (1998). Principles of geographical information systems. Oxford: Oxford University Press.
- Goodchild, M. F., Parks, B. O., & Steyaert, L. T. (1993). Environmental modeling with GIS. New York: Oxford University Press.
- Tomlin, C. D. (1990). Geographic information systems and cartographic modeling. Englewood Cliffs, NJ: Prentice-Hall. [**map algebra**]

Some categorizations

- Raster operations:
  - **local**: at a pixel, e.g., transformations
  - **focal**: around a pixel, in its neighbourhood, e.g., filter
  - **zonal**: pixel in some map unit or 'zone', e.g. mean value of all pixels in the map unit
  - **global**: all pixels, e.g., distance from a source

- Conceptual:
  - Extract attributes at a location
  - spread attributes or some function over the map, e.g., buffer
  - overlay: combine maps

Organization of tools in QGIS “toolboxes”



Tools are organized by source (plugin) and then categorized conceptually

Steps in modelling

These apply to any sort of model; the terminology here is mostly from empirical-statistical ("regression") models.

1. Selecting a **functional form**, i.e. the model to be fitted;
   - may try several forms; but these should be "reasonable", based on possible mechanisms
2. Determining the **parameters** of the model; this is called **calibration** or **parameter estimation**;
3. Determining how well the model describes reality; this is called **validation**.
4. **Criticising** (re-examining) the assumptions and possibly re-cycling.
   - may lead to a modified or completely different **model form**

Examples of functional forms

Grain yield of a cereal crop as affected by N fertilizer:

1. **linear**: one unit of fertilizer is β units of grain yield, throughout the range;

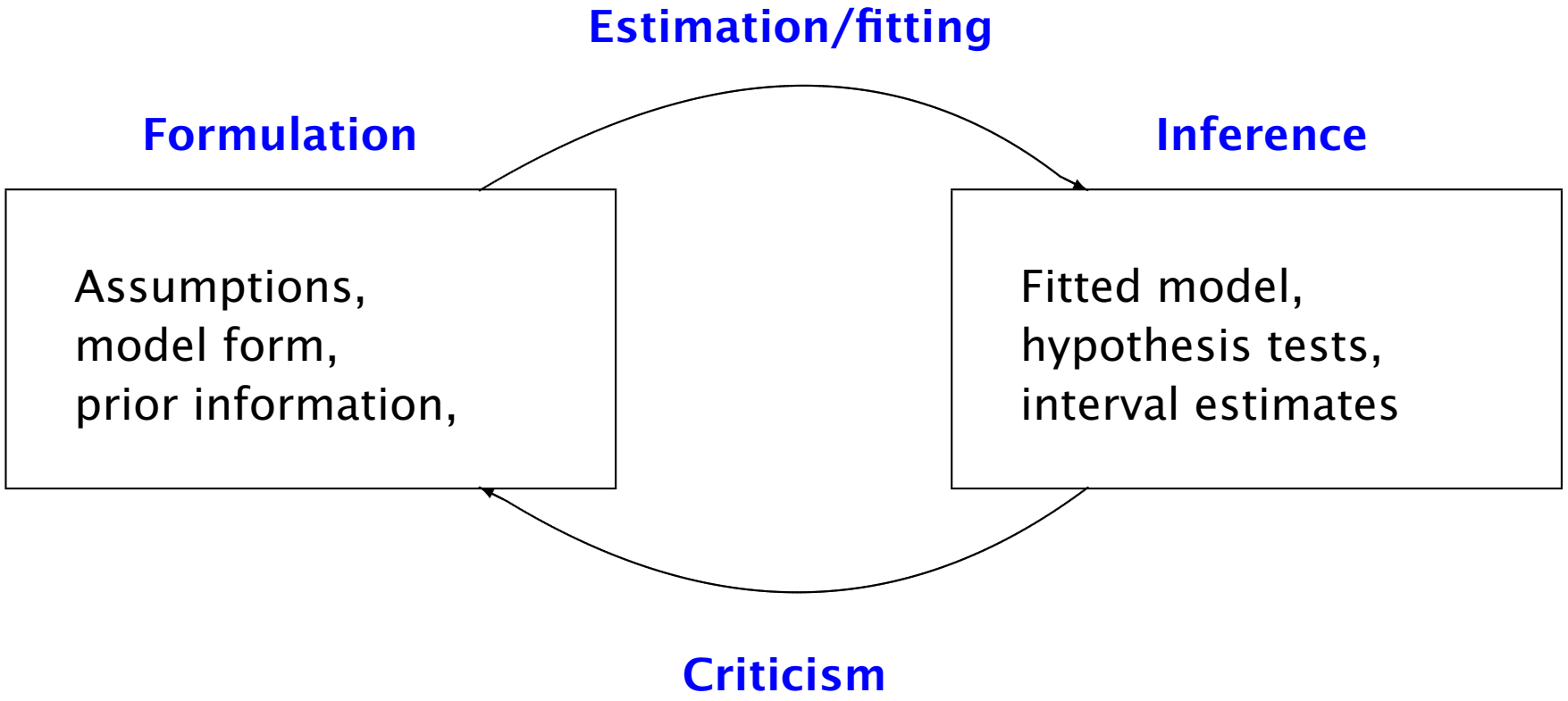
2. **linear response with threshold**: same till λ units of N, then reaches a plateau;

3. **quadratic**: one unit of fertilizer is $\beta_1 + \beta_2{}^2$ units of grain yield, throughout the range; β_2 is negative so after a certain point yield decreases;

4. **negative exponential**: yield increases asymptotically to some limit μ at some effective range ρ.

- The Greek letters indicate **parameters** that must be fit by **calibration**
- All of these have a **plausible physical basis** within a **range of applicability**

The modelling paradigm



– after Cook & Weisberg (1982) *Residuals and influence in regression*

Note **criticism** of the **assumptions**, especially **model form**.

Structure vs. noise

- Observations = f(Structure, Noise)
- Observations = f(model, unexplained variation)

Observations are a subset of **Reality**, so:

- Reality = f(Structure, Noise)
- Reality = f(deterministic **processes**, **random** variation)

The aim is to match our **model** with the true **deterministic process** . . .

. . . and match our estimate of the **noise** with the actual **random variation**.

It is equally an error to model the noise (**overfit** the model) as to not model the process (**underfit** the model).

Evidence that a model is suitable

Two levels of evidence:

1. **external** to the model:
   (a) what is known or suspected about the **process** that gave rise to the data
   (b) this is the connection to the **reality** that the model is trying to explain or summarise;
   (c) how well the model fits further data from the same population: success of **validation** against an independent dataset

2. **internal**: from the model itself:
   (a) how well the model fits the data (success of **calibration**);
   (b) how well the fitted model meets the **assumptions** of that functional form (e.g. examination of regression diagnostics).

Example of a spatial modelling exercise

- **Problem**: soil contamination by heavy metals in flood plain of the Maas (Meuse) River near Stein (L), Netherlands

- **Objectives**:

1. determine where metals came from (**explanation**)
   - original sediments?
   - recent atmospheric deposition from industry?
   - sediments deposited by floods from upstream (B, F)?
2. map sediment concentrations over an area (**prediction**)

These objectives are not mutually exclusive! In fact, better explanation often leads to better predictive models.

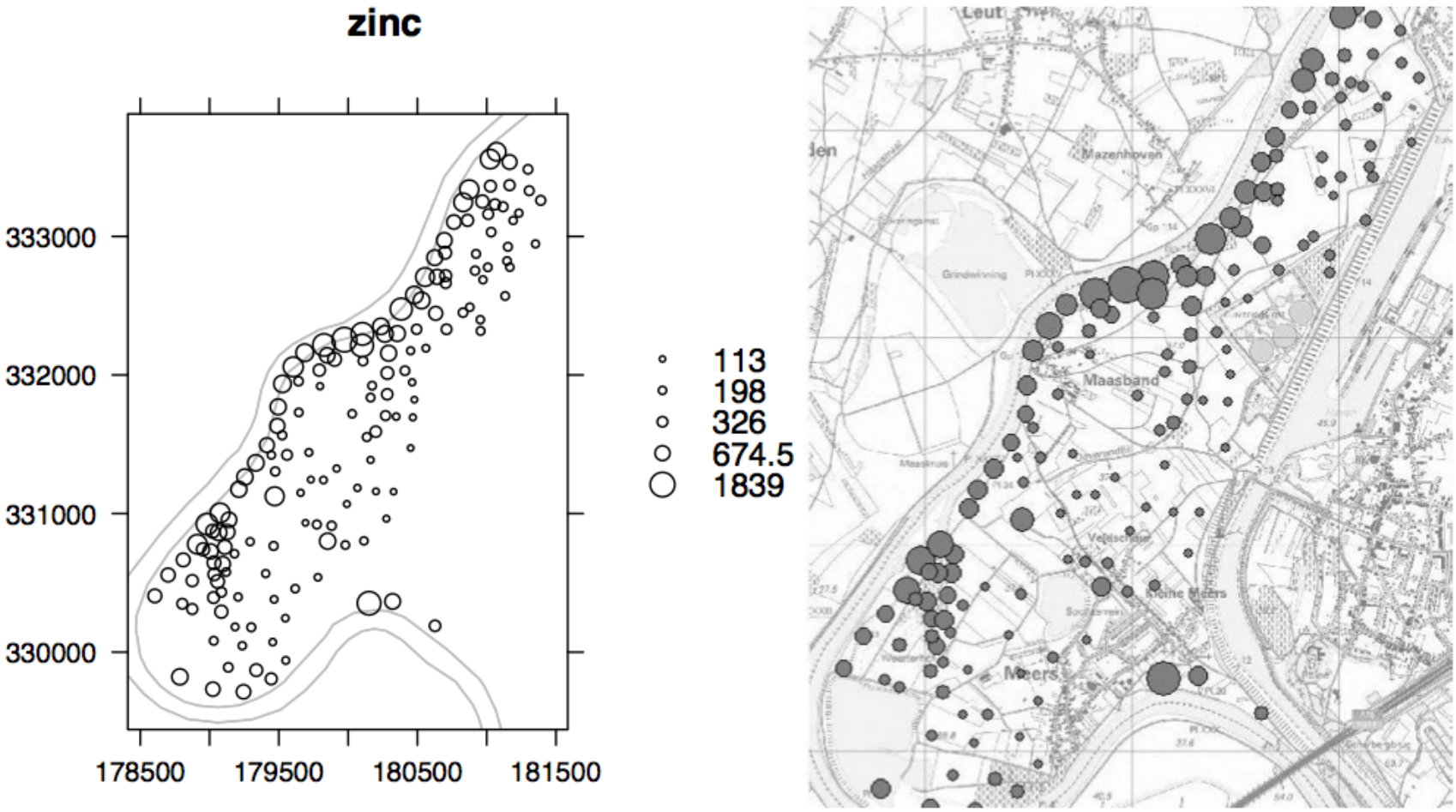


Fig. 5.2: Meuse data set and values of zinc (ppm): visualized in R (left), and in SAGA GIS (right).

Source: Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*

Chosen model forms

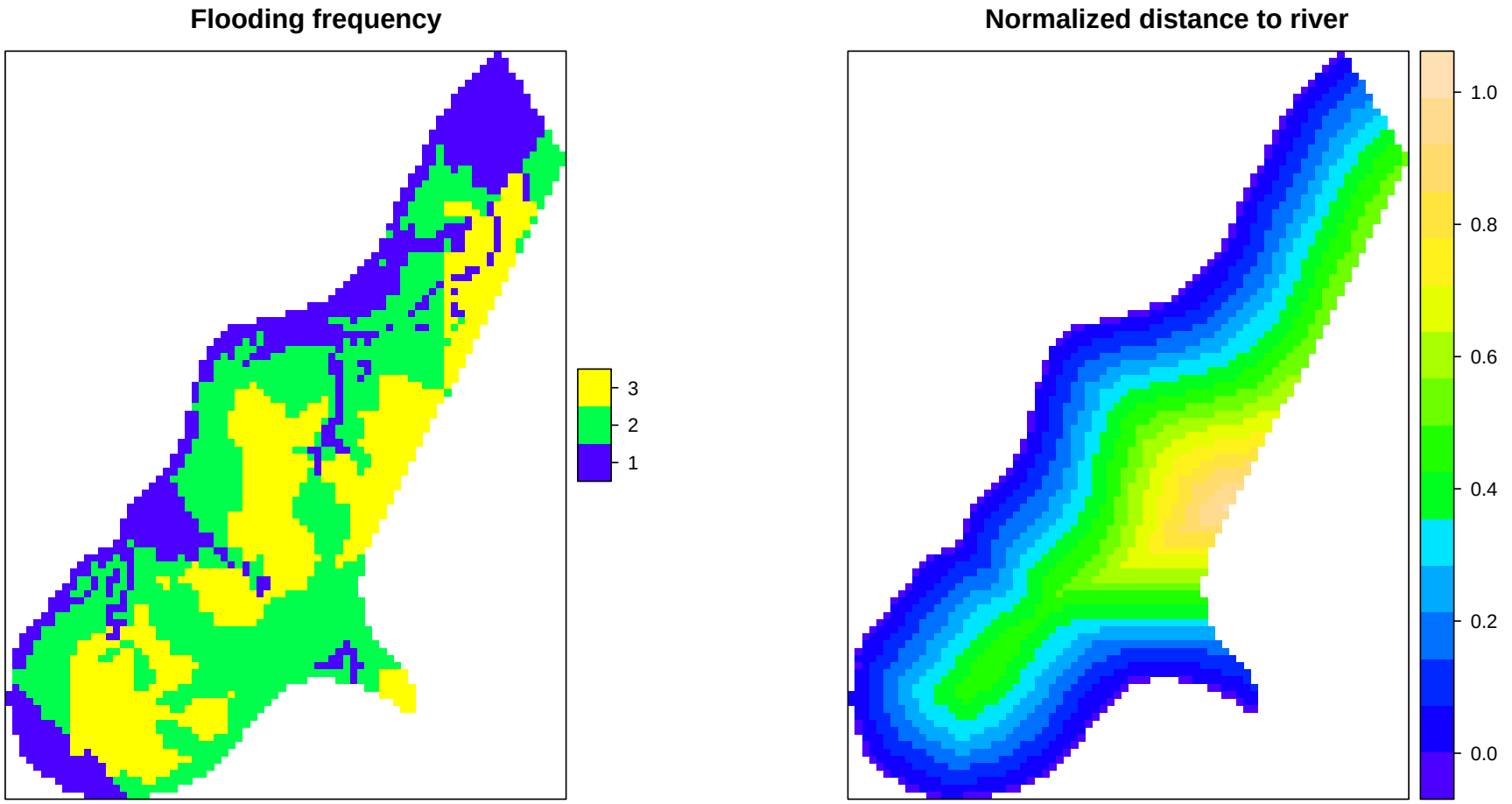
(1) **Ordinary kriging (OK)**: local spatial dependence (covariance) of **soil property**

- theory: metal is the result of a spatially-autocorrelated random process
- ignores obvious geographic co-variates – a convenient mathematical fiction

(2) **Regression kriging (RK)**: combines:

1. **Feature space** model: metal vs. covariables: flooding frequency, distance from river, soil organic matter
   - using multiple linear regression, possibly with some transformations
   - e.g., `log(zinc) ~ ffreq * dist`
   - These predictors expected from theory/previous studies
2. **Geographic space** model: local spatial dependence (covariance) of **residuals** from feature-space model
   - using OK of the residuals – variation not explained by feature-space model

Feature-space predictors



These *are* distributed in geographic space; and the normalized distance is derived by a geographic-space operation, so in that sense this is a “spatial” model, even though evaluated with a feature-space model.

Modelling results – feature space (1)

```
R> summary(m <- lm(log(zinc) ~ ffreq*dist, data=meuse))

Residuals:
    Min      1Q  Median      3Q     Max
-0.9439 -0.3107 -0.0058  0.2488  1.6676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7692     0.0702   96.44  < 2e-16
ffreq2       -0.7664     0.1319   -5.81  3.6e-08
ffreq3       -0.5579     0.1954   -2.86  0.00491
dist         -3.0909     0.2885  -10.71  < 2e-16
ffreq2:dist   1.4180     0.4023    3.52  0.00056
ffreq3:dist   0.8281     0.6384    1.30  0.19657

Residual standard error: 0.436 on 149 degrees of freedom
Multiple R-squared: 0.647,Adjusted R-squared: 0.635
```

Interpretation:

(1) reasonably successful model (almost 2/3 variance explained);

(2) Flood frequency 2 & 3 significantly lower levels than 1;

(3) concentration decreases with distance from river;

(4) this effect is less for areas with flood frequency 2.

Modelling results – feature space (2)

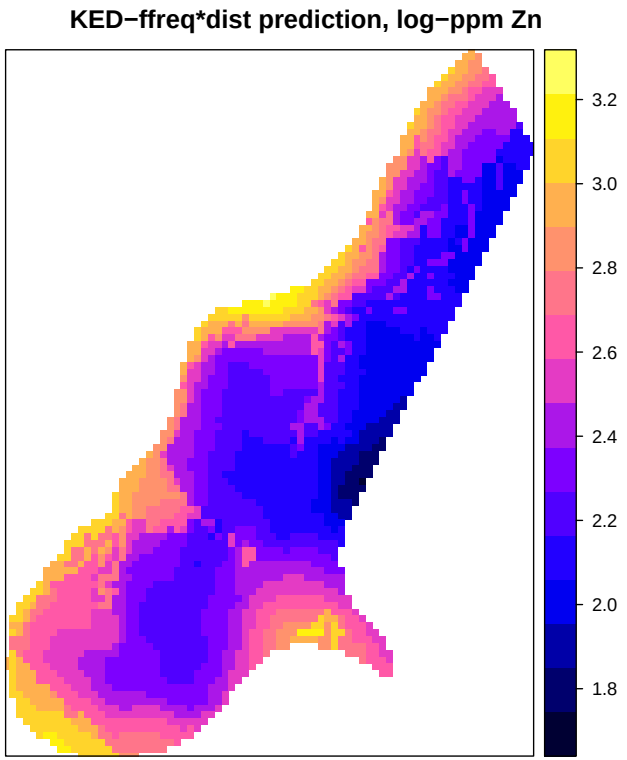
After controlling for distance to river and flooding frequency, is there an effect of **soil organic matter**?

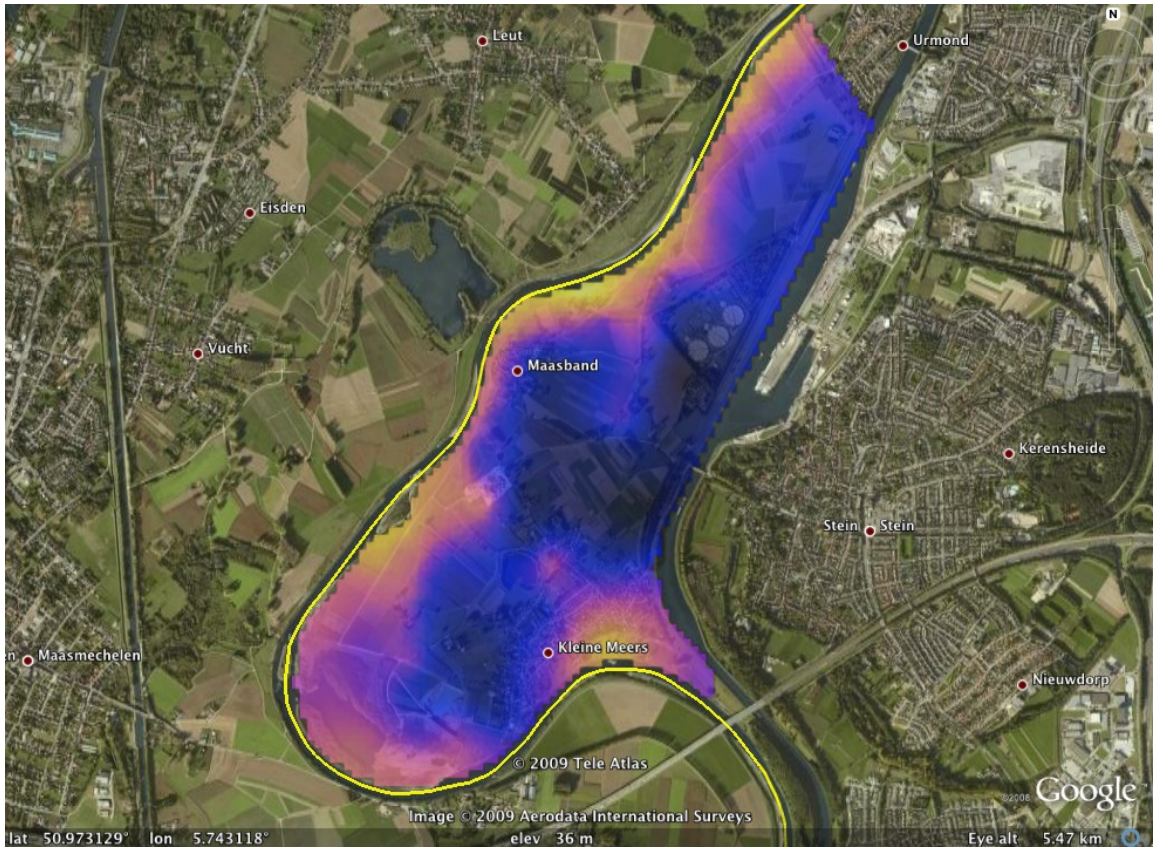
```
R> summary(m.om <- lm(residuals(m) ~ om, data=meuse))

Residuals:
    Min      1Q  Median      3Q     Max
-0.7765 -0.3263 -0.0283  0.2725  1.6762

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.25022    0.07834   -3.19  0.00171
om           0.03501    0.00953    3.67  0.00033

Multiple R-squared: 0.0821,Adjusted R-squared: 0.076
```

Interpretation: Almost no relation between these soil constituents, once the common (presumed) underlying variables are accounted for.

Note: we rely on physical principles to assert that river flooding affects organic matter content, not the other way around!

Modelling results – as prediction maps





Regression kriging

ordinary kriging

Clear effect of flood class and distance maps

One smooth surface

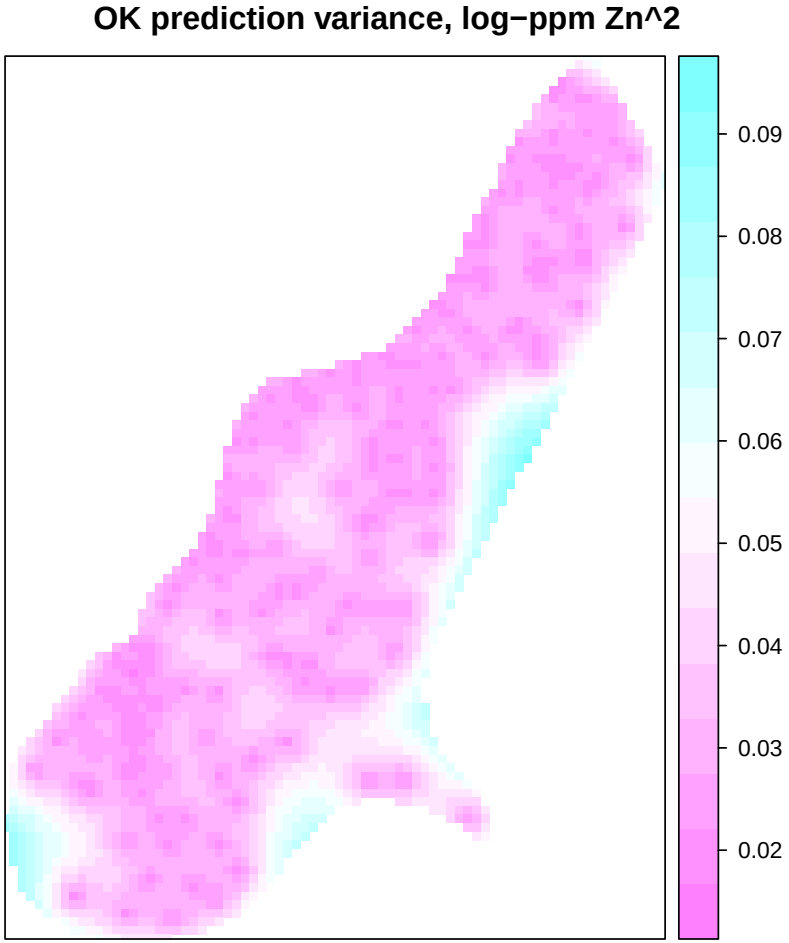
Modelling results – inferences

What does the model **imply** about the processes, i.e., what can we **infer**?

1. Strong relation between metal concentration and (1) flood frequency; (2) interaction of this with distance to river
   - **inference**: metals are from upstream industry; flood waters spread them on this land
2. No relation with organic matter
   - **inference**: little or no reaction with soil in short term
3. Strong spatial dependence of residuals
   - **inference**: local "hot" and "cold" spots of deposition, perhaps from specific flood events in different places?

These inferences are not **proven**, rather, the evidence allows us to argue their **plausibility** with respect to **competing hypotheses**

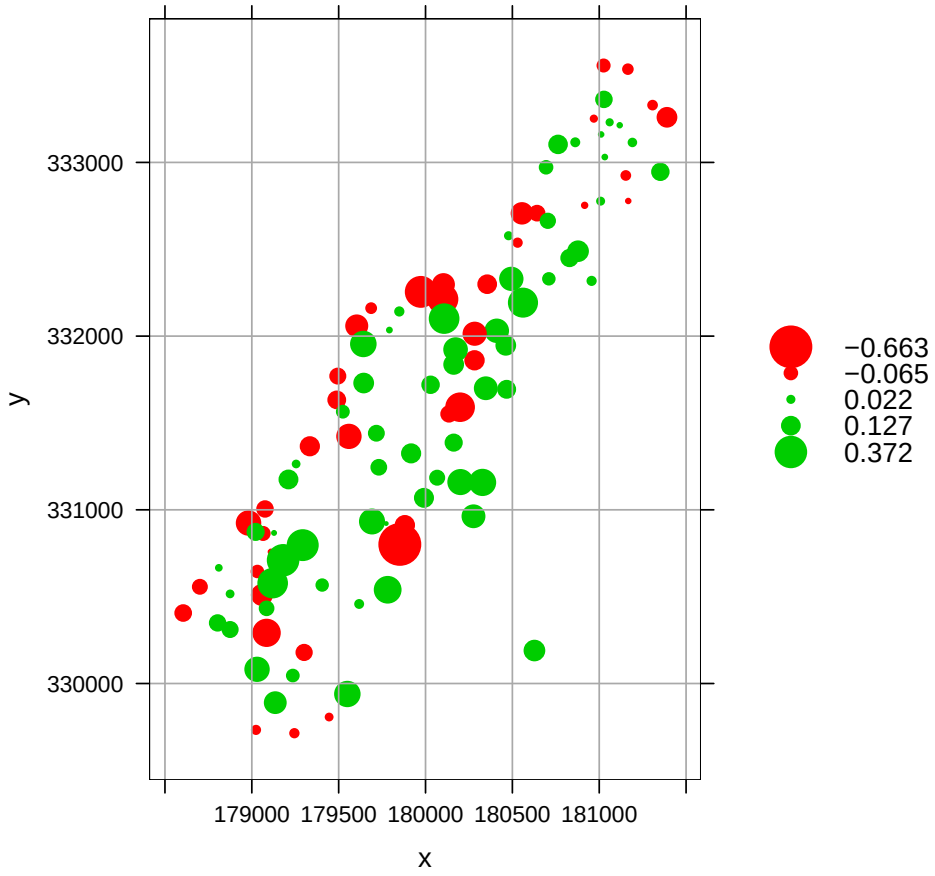
Internal model quality: kriging prediction variance



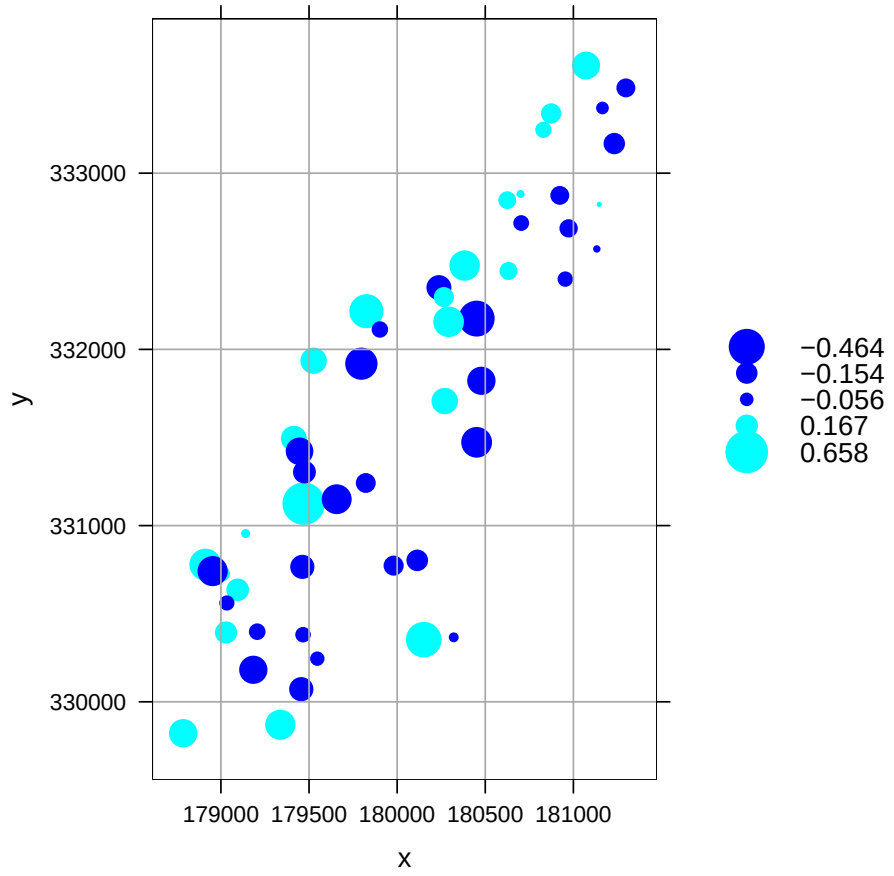
The variance is produced along with the prediction.

External model quality: kriging prediction validation

OK validation errors at undersampled points, log10(Pb)



OK cross−validation errors, log10(Pb)



Prediction errors, applying the model at known points, not used in the prediction.