Applied geostatistics

Lecture 3 – Modelling spatial structure from point samples

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

October 16, 2010

Copyright © 2007–2010 University of Twente, Faculty ITC.
All rights reserved. Reproduction and dissemination of the work as a whole (not parts) freely permitted if this original copyright notice is included. Sale or placement on a web site where payment must be made to access this document is strictly prohibited. To adapt or translate please contact the author (`http://www.itc.nl/personal/rossiter`).

ITC

Topics for this lecture

1. Trend surfaces
2. Theory of random fields
3. Models of spatial covariance
4. Variogram analysis; variogram model fitting

Commentary

We often have a set of **point observations** (with some spatial support around the "point"). From these we usually want to:

1. **identify** the **nature** of any **spatial structure** of one or more feature-space **attributes** measured at each observation point;
2. **infer** what **spatial processes** caused the observed structure;
3. **interpolate** (predict) the attribute at **unsampled locations**, often on a regular grid to cover an entire study area.

We divide this into two **types of processes**:

1. **Regional**, also called **global** by contrast to the next; the process operates over an entire region;
2. **Local**: the process operates over some **neighbourhood** ("range") beyond which there is no influence.

Topic 1: Trend surfaces

Regional processes operate over an entire region; that is, **one mathematical model** is used to describe such a process.

The most common regional model is the **trend surface**.

Regional models

- **Theory**: The attribute value depends on **relative geographic position** within a spatial field

- That is, the **process** that affected the attribute value operates over the whole region; it has one cause

  * **Example**: thickness of a layer of volcanic ash over a buried soil: one source of ash (a volcano), spread by prevailing winds; therefore **one process**

- Since there is only one process ("global"), **all** sample points are used to fit **the** model.

  * Note: Sometimes the same model form is fitted **piecewise** over the region, e.g. in moving trend surfaces or splines; but we will ignore that for now.

To check your understanding . . .

Q1 : *Describe a process in your field of study that would be expected to show this behaviour, i.e. the attribute value depends on relative geographic position within a spatial field* *Jump to A1* •

Commentary

There are many mathematical forms for a regional trend. We will see the most common, namely the **polynomial trend surface**.

However, you should be aware that there are other forms to model surfaces.

Trend surfaces

- A **mathematical model** of a **regional trend**
- The value of a variable at each point depends only on its **coördinates** and parameters of a fitted surface
- This is modelled with a **smooth function of position**; in 2D this is:

$$z = f(x, y)$$

where z is the attribute value and x, y are the two coördinates, e.g. UTM E and N.

- The function f is generally **differentiable** at least twice, to give slopes and curvatures.
- When this function is computed on a **regular grid**, the result is called a **trend surface**.

Polynomial trend surfaces

The most common trend surface functions are **polynomials** in the coördinates. These have some good mathematical properties:

- They are **differentiable** to any degree (of course, beyond the degree of the polynomial the derivative is zero).
- They are easy to **fit** using some variety of **least-squares**.

The simplest trend is the **plane**, also known as the **first order** trend:

$$z = \beta_0 + \beta_x x + \beta_y y$$

General polynomial trend surfaces

- General form: a surface of order p:

$$f(x,y) = \sum_{r+s \leq p} \beta_{r,s} x^r y^s$$

- Example: full 2nd order:

$$z = \beta_0 + \beta_{1,0}x + \beta_{0,1}y + \beta_{2,0}x^2 + \beta_{0,2}y^2 + \beta_{1,1}xy$$

- Order should be suggested by the process
  - One-way trend: 1st order (plane)
  - Dome or depression: 2nd order
  - Folded structure: Higher orders, depending on number of inflection points

- Higher orders always give better fits to the data, but beware that you may be fitting **noise**, not **structure**; use the adjusted R^2 or similar to evaluate.

Fitting trend surfaces

- The trend surface is modelled by **linear regression** with **coördinates as the predictor variables**, using data from all sample points.
- In **ordinary** linear regression, all samples participate **equally** in the prediction
- The **goodness of fit** of the trend surface **to the sample** is measured by the **residual sum of squares**, or, equivalently, the R^2 of the regression.
- The same **cautions** apply as in feature-space regression analysis!
- **Ordinary Least Squares** (OLS) is commonly used to compute the regression coefficients.
  - Note: OLS is not really correct, since it ignores possible **correlation among closely-spaced samples**; better is **Generalised Least Squares** (GLS); that is an advanced topic, see Lesson 5, "Regression Kriging"

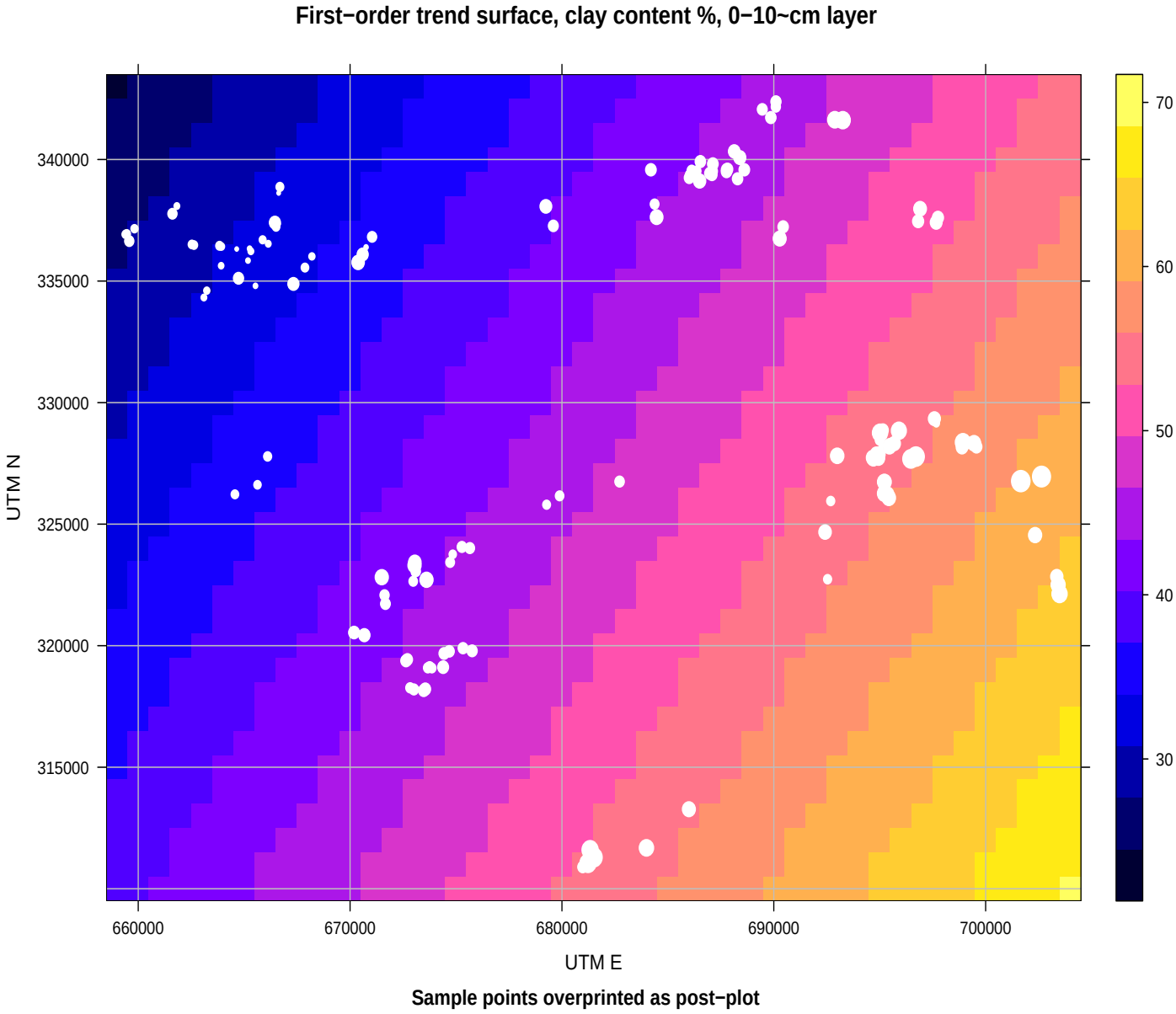
Commentary

In the previous lecture and exercise ("visualization") we computed and displayed first- and second-order trend surfaces for subsoil clay content in a region of southern Cameroon.

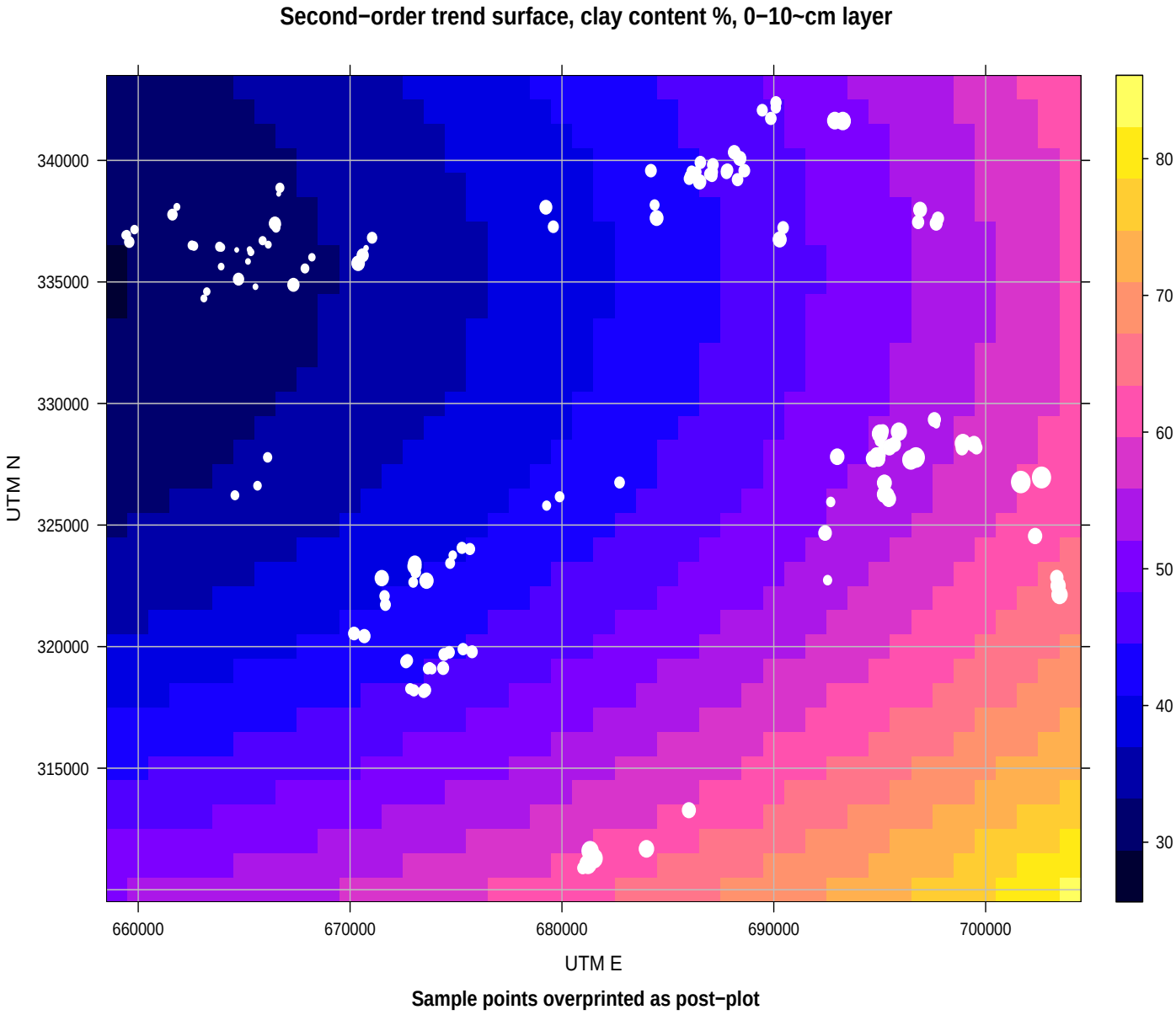
However, we didn't see how to model the trend with linear regression, nor how to evaluate the model.

To refresh your memory, the next two slides are the two computed surfaces, then we examine the models behind them, and how to interpret the models and fit.

First-order trend surface



Second-order trend surface



Computing a trend surface model by ordinary least squares regression

We use the Cameroon soils data of Yemefack *et al.* (Geoderma, 125:117)

Each observation has two UTM coördinates and two attributes:

```
      UTM_E  UTM_N clay35 pH35
1 702638 326959     78 4.80
2 701659 326772     80 4.40
3 703488 322133     66 4.20
4 703421 322508     61 4.54
5 703358 322846     53 4.40
6 702334 324551     57 4.56
...
```

We use the R `lm` ("linear models") method; all statistical packages have an equivalent comment to compute a least-squares fit:

```
ts1 <- lm(clay35 ~ UTM_E + UTM_N, data=tcp)
```

The **model formula** `clay35 ~ UTM_E + UTM_N` specifies the functional form, in this case a dependency on the two coördinates.

Computing a trend surface (2)

The result of this computation is a **fitted linear regression** of the **attribute** (here, clay content) on the **predictors**, which are the grid **coördinates**.

R gives us the following summary table; other statistics packages have something quite similar:

```
Residuals:
    Min      1Q  Median      3Q     Max
-31.601  -5.106  -0.363   3.607  20.467

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.50e+02   5.19e+01   -4.83  3.5e-06
UTM_E        6.51e-04   5.97e-05   10.91  < 2e-16
UTM_N       -4.50e-04   9.24e-05   -4.88  2.8e-06

Multiple R-Squared: 0.506,        Adjusted R-squared: 0.499
```

Interpretation of the linear model output

In the summary of the linear model we see:

Residuals **Lack of fit** of individual observations; here from -31 to $+20$ percent clay;

Coefficients **Multipliers** of each term in the polynomial; here for every meter East, the clay content increases by 0.00065%, i.e. 0.65% per km

Adjusted R-squared **Proportion of variance explained** by the model, here about 50%.

Recall: The "adjusted" R^2 decreases the apparent R^2, computed from the Analysis of Variance (ANOVA) table, to account for the number of predictive factors:

$$R^2_{\text{adj}} \equiv 1 - [(n-1)/(n-p) * (1-R^2)]$$

where n is the number of observation and p is the number of coefficients.

To check your understanding . . .

Q2 : *What is the change in clay content for each km towards the North?* *Jump to A2* •

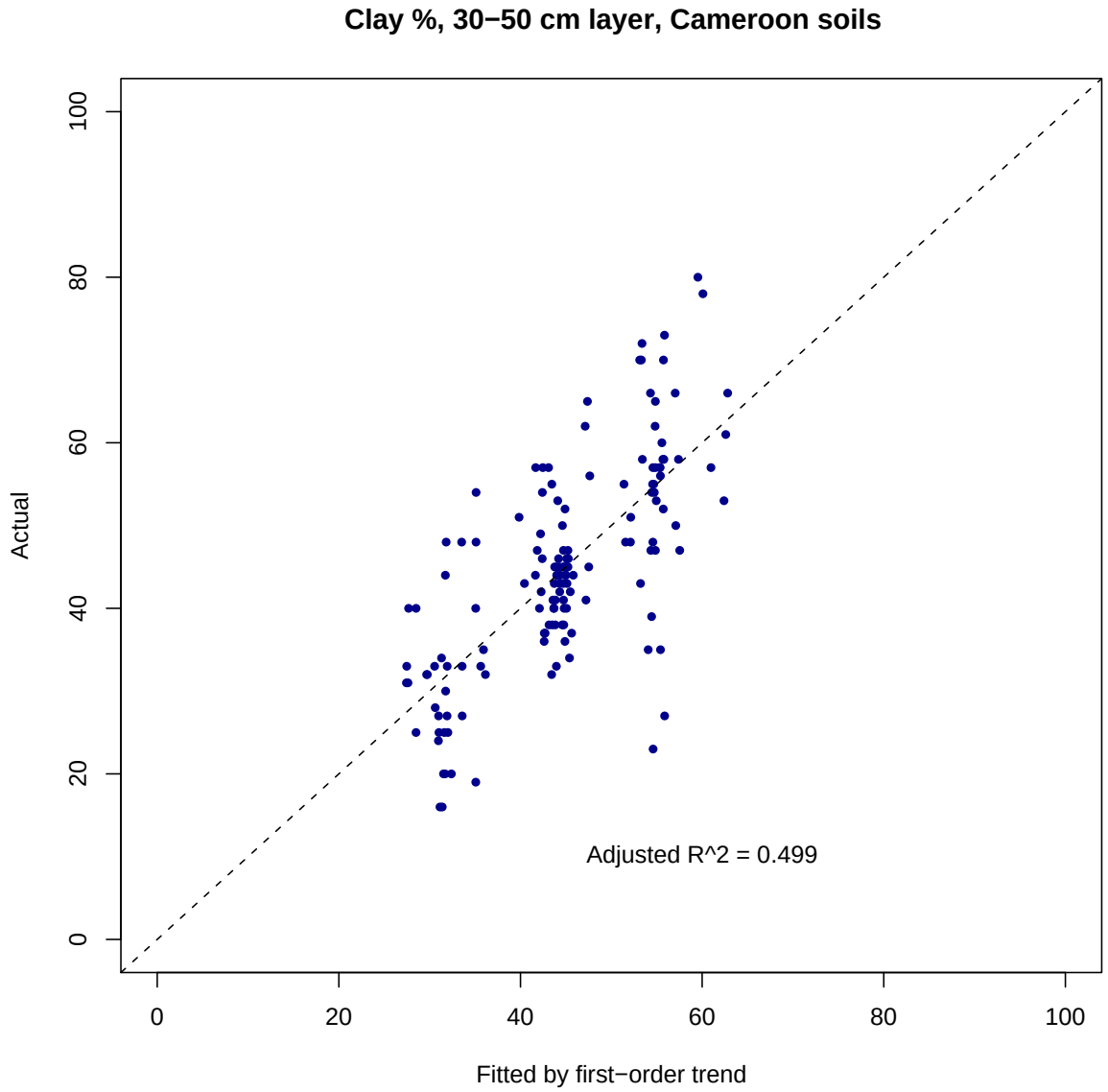
Q3 : *What proportion of the residuals are between* -5.106 *and* $+3.607$ *% clay?* *Jump to A3* •

Evaluating a trend surface fit (1)

The same tools available for any linear regression are available here:

- **Numerical** measures:
  - **Adjusted** R^2: Proportion of variance explained by the model
  - **Residuals**
- Graphical method 1: "1:1" plots: **Fitted** vs. **observed** values
  - Ideally all points should fall on the 1:1 line
  - If not all on line, equally-spread on both sides
  - If there is a "curve" apparent, probably need a higher-order polynomial

Fitted vs. observed (1:1) plot



To check your understanding . . .

Q4 : *How well do the fitted values match the observed values? What are the deviations from the 1:1 line?*

Jump to A4 •

Evaluating a trend surface fit (2)

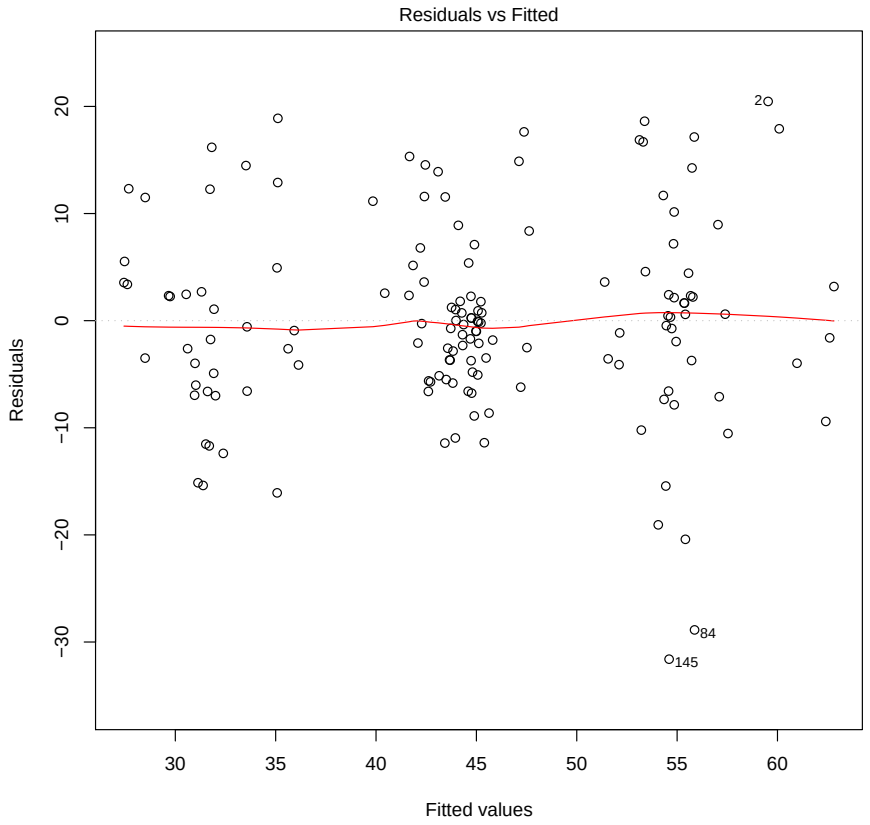
- Graphical method 2: **Diagnostic plots of residuals**

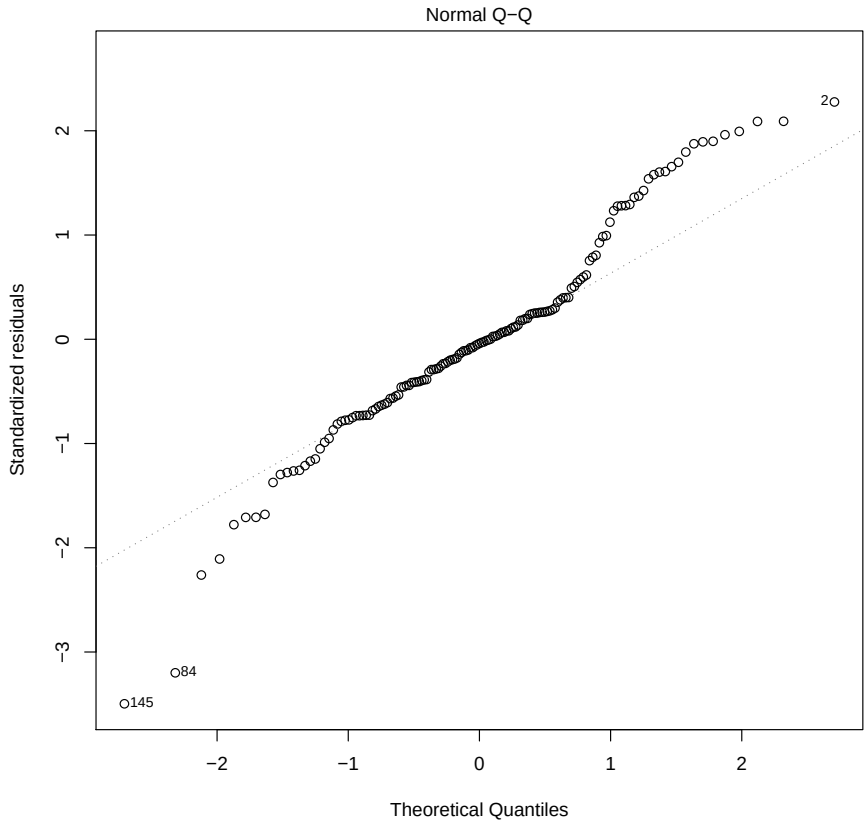
- These show whether the **residuals** satisfy the model assumption; in particular, they should be **normally distributed**

  1. **Fitted** values vs. **residuals**: residuals should show no pattern
     - if they are more variable in some part of the range then you have **heteroscedascity** and need a **variance-stabilizing transformation**
     - if they are mostly on one side of the 0-axis in some range you probably need a higher-order polynomial or non-linear model.
  2. **Quantile-quantile (Q-Q) plot** of **residuals**: compares theoretical quantiles of a normal distribution with actual quantiles
     - all points should be on the line
     - should be denser in the centre (median) and spread out at the two ends.

- See a good regression text for more information, e.g. Fox, J. (1997). Applied regression, linear models, and related methods. Newbury Park: Sage.

Diagnostic plots from linear regression





To check your understanding . . .

Q5 : *Is there any pattern of the residuals vs. fitted values?* *Jump to A5* •

Q6 : *Are the residuals normally-distributed?* *Jump to A6* •

Q7 : *Which observations are marked by the diagnostic procedure as being outliers in this linear model fit?* *Jump to A7* •

Computing a higher-order trend surface

Some trends require a **higher order** than a simple plane. They are fit in the same way as the first-order surface, but include higher powers of the coördinates as predictors.

In the present example it is clear from the surface/post-plot and the regression diagnostics that a first-order surface was not satisfactory. So we compute with the square of the coördinates and their cross-product included in the model, and get the following summary:

```
ts2 <- lm(clay35 ~ I(UTM_E^2) + I(UTM_N^2) + I(UTM_E*UTM_N) + UTM_E + UTM_N, data=tcp)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-29.48  -5.00  -0.56   4.21  20.83
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.18e+04   6.74e+03    1.75   0.0822 .
I(UTM_E^2)        1.27e-08   8.44e-09    1.50   0.1349
I(UTM_N^2)        3.58e-08   1.26e-08    2.83   0.0053 **
I(UTM_E * UTM_N)  1.02e-08   1.33e-08    0.77   0.4425
UTM_E            -2.00e-02   1.48e-02   -1.35   0.1777
UTM_N            -3.10e-02   1.24e-02   -2.50   0.0135 *

Multiple R-Squared: 0.534,        Adjusted R-squared: 0.518
```

Interpretation of the polynomial model output

Residuals Lack of fit of individual observations; here from -29.5 to $+20.8$ percent clay; about the same as for the first-order surface

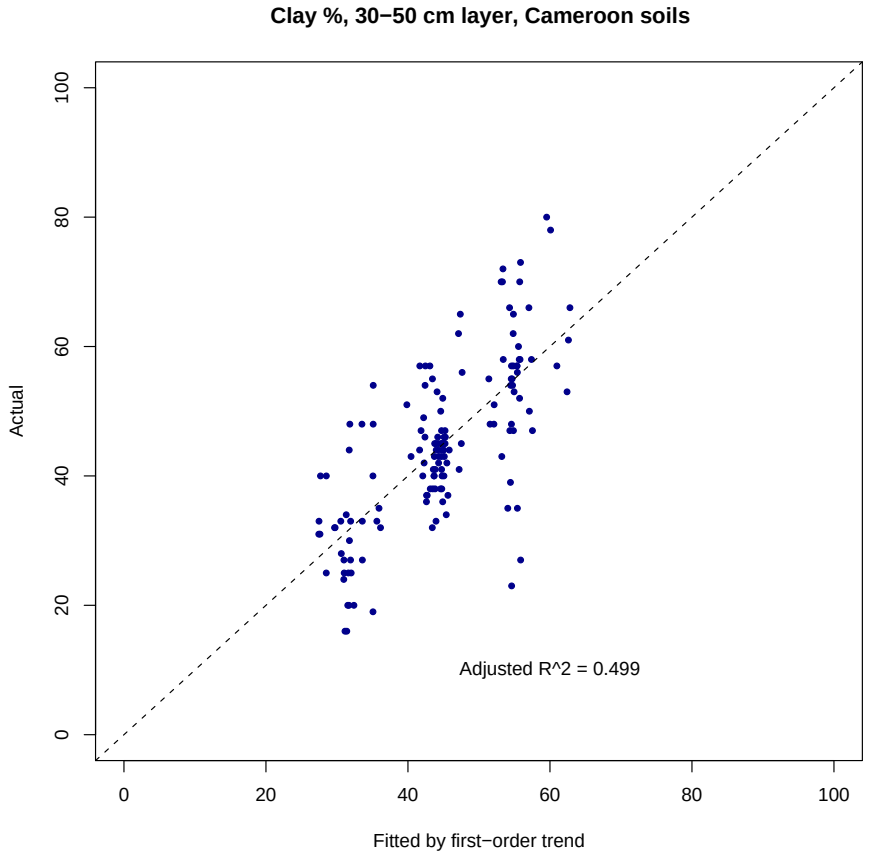
Coefficients Multipliers of each term in the polynomial, but now there are four: the coördinates and their squares, and of course the intercept.

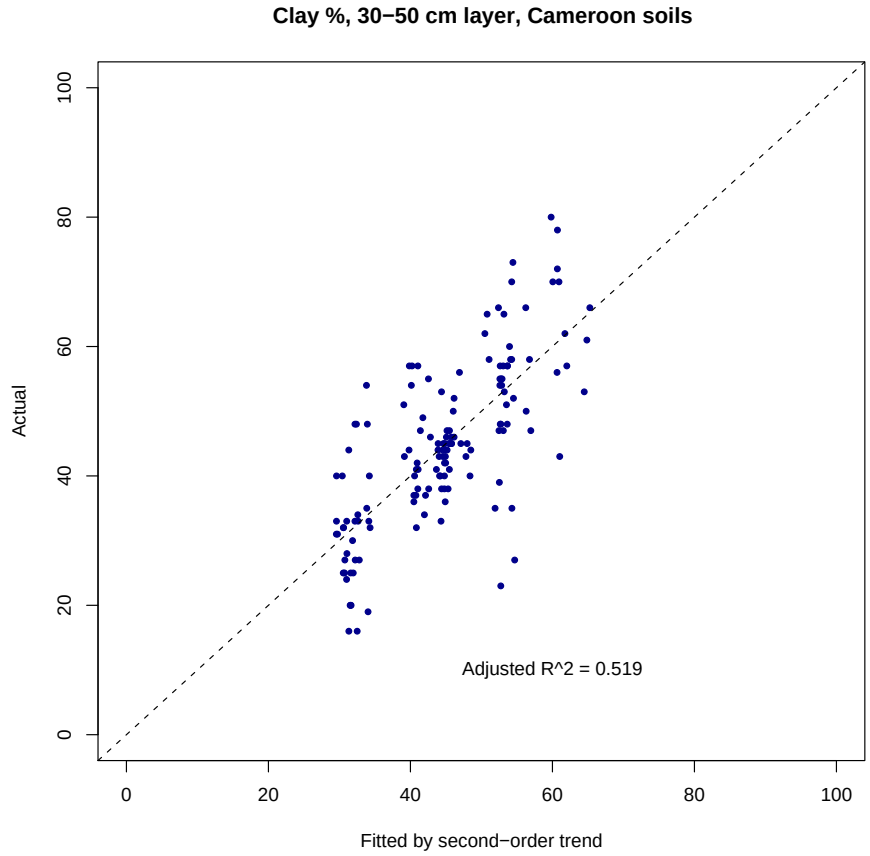
Statistical significance of coefficients The listed `Pr(>|t|)` gives the probability that the coefficient is in fact 0 (the null hypothesis), i.e. that it contributes nothing to the model. In this example the cross-product of the coördinates, i.e. `I(UTM_E * UTM_N)` is almost surely not necessary and the model can be re-fit without it.

Adjusted R-squared Proportion of variance explained by the model, here about 52%, a slight improvement over the first-order surface

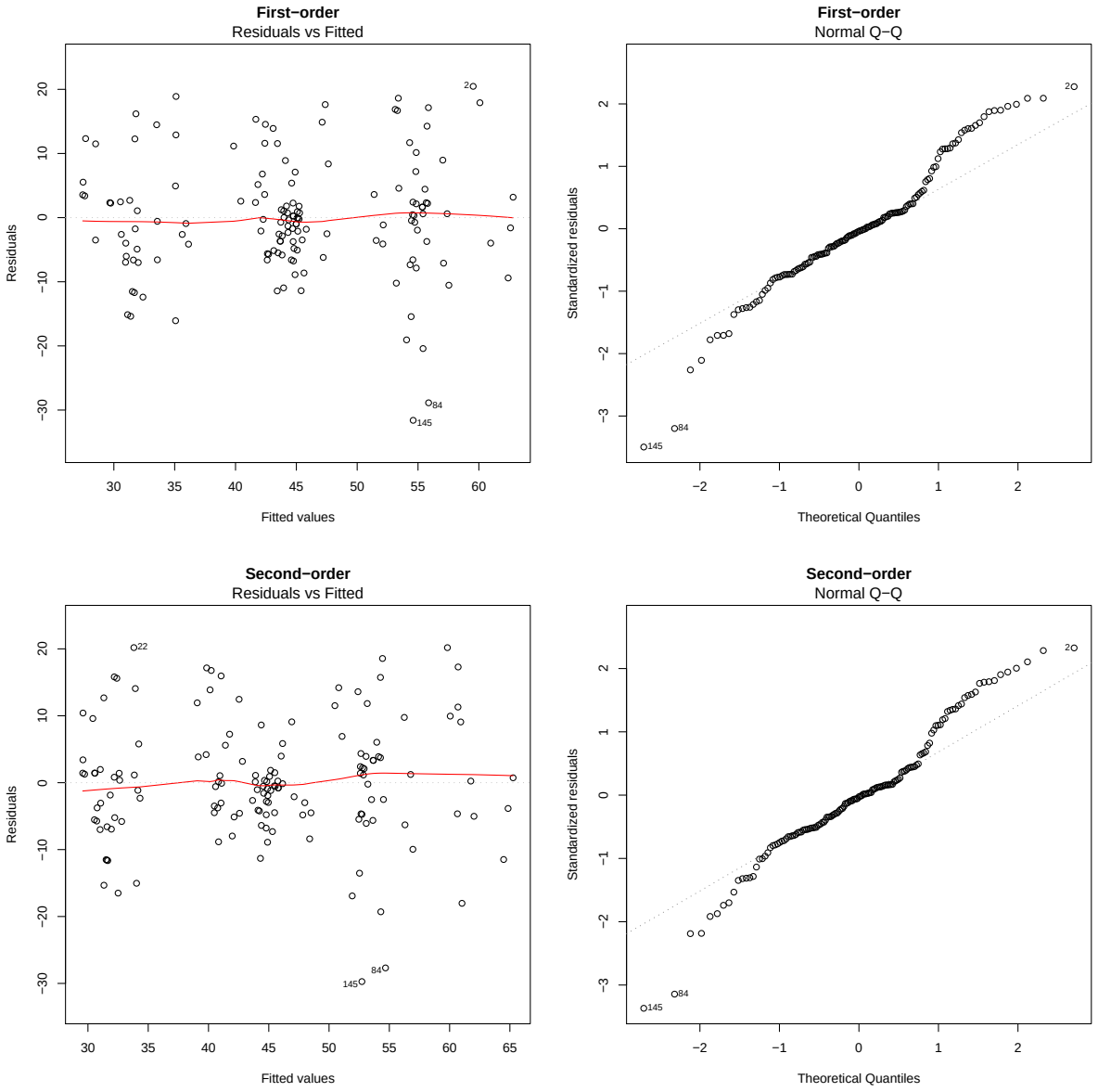
We evaluate the fit with the same diagnostics as in the first-order model.

Fitted vs. observed (1:1) plots





Diagnostic plots from first- and second-order linear regressions



To check your understanding . . .

Q8 : *How much improvement is there with the second-order, compared to the first-order, trend surface? Comment on the adjusted R^2, the fitted vs. observed values, and the diagnostic plots of the residuals. Jump to A8* •

Commentary

In summary, there are some **processes** that operate over an **entire region** of interest. These can lead to **regional trends**, where an attribute can be modelled as a **function of the geographic coördinates**.

A common functional form is the **polynomial trend surface**, which is fit to the data by some form of **regression**, often **ordinary least squares**.

However, as the example in this section showed, the regional trend often does not explain all the variation.

We now look at the other side, namely processes that operate only **locally**. Later we will see how to combine the two types of spatial structure.

Topic 2: Theory of random fields

Now we turn to **local spatial dependence**. The idea here is that there is a **local process** that causes nearby points to be "similar". We will see how to quantify this and use it in modelling and prediction.

But first we need some **theoretical background**: a brief explanation the theory underlying optimal geostatistical estimation by kriging.

The presentation is based on R. Webster and M. Oliver, 2001 *Geostatistics for environmental scientists*, Chichester etc.: John Wiley& Sons, Ltd.

Commentary

This is difficult material, mainly because it is **abstract** and indeed quite strange at first glance. Some modellers do not accept it at all! Our position is that this **model** of an essentially **unobservable** hypothetical process has proven to be very **useful** in practice.

The important point here is that we believe that **observed values** in space are the result of some **spatially-correlated process** (to be explained) which we can **infer** from the observations.

But be aware: as with many models, it is **impossible to verify** the underlying **assumptions**. If you feel uncomfortable with this, you are not alone!

Modelling spatial variation as a random process

Notation: A point in space of any dimension is symbolized by a **bold-face** letter, e.g. $\mathbf{x}$.

- **Key idea**: The observed attribute values are only **one of many possible realisations** of a **random process** (also called a "stochastic" process)
- This random processes is **spatially auto-correlated** (to be explained below), so that attribute values are somewhat **dependent**.
- At each point $\mathbf{x}$, an observed value z is one possiblility of a **random variable** $Z(\mathbf{x})$
- There is **only one reality** (which is sampled), but it is one **realisation of a process** that **could have** produced many realities. μ and variance σ^2 etc.
- CDF: for a given value z: $F\{Z(\mathbf{x};z)\} = \Pr[Z(\mathbf{x} \leq z_c)]$
- the probability $\Pr$ governs the random process; this is where we can model **spatial dependence**

Random functions

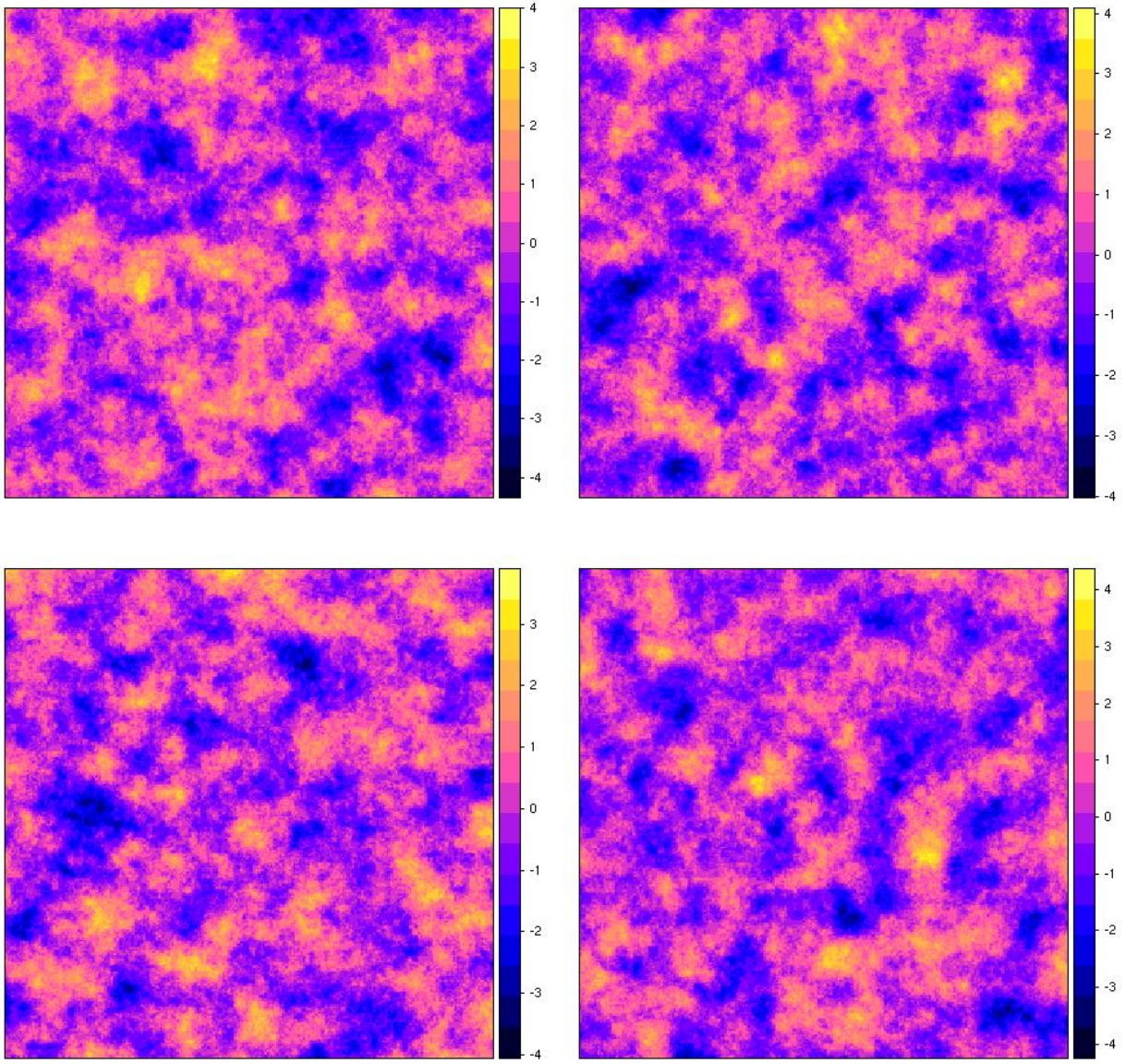
- Each point has its **own** random process, but these all have the **same form** (same kind of randomness)
- However, there may be **spatial dependence** among points, which are therefore *not* independent
- As a whole, they make up a **stochastic process** over the whole field R
- i.e. the observed values are assumed to result from some random process but one that **respects certain restrictions**, in particular spatial dependence
- The set of observed values: $Z = \{Z(\mathbf{x}), \forall \mathbf{x} \in R\}$ is called a **regionalized variable**
- This variable is **doubly infinite**: (1) number of points; (2) possible values at each point

Commentary

The above is a difficult point for most people to grasp. It's always easier to **visualise**, so the next slide shows four **different realizations** of the **same random field**.

(For later reference: this is a 256 x 256 grid, with a Spherical model of spatial dependence with range 25)

Four realizations of the same random field



To check your understanding . . .

Comparing the four realizations:

Q9 : *How similar are the* **attribute values** *at any* **specific location***?* *Jump to A9* •

Q10 : *How similar are the* **patterns***? Compare the patch sizes, sharpness of transitions between patches, etc.* *Jump to A10* •

Commentary

So there is a **regionalized variable**, i.e. where every point is the outcome of some random process with spatial dependence.

But there is a serious problem: there is only one **realization**; it's as if we had to model a trend surface from one point!

We can get around this problem by making some **assumptions**, collectively known as **stationarity**.

First-order stationarity

- **Problem**: We have no way to estimate the expected values of the random process at each location $\mu(\mathbf{x})$ ...
- ...since we only have **one realisation** (what we actually measure), rather than the whole set of realisations that could have been produced by the random process.
- **Solution**:**assume that the expected values at all locations in the field are the same**:

$$E[Z(\mathbf{x}_i)] = \mu, \forall \mathbf{x}_i \in R$$

- This is called **first-order stationarity** of the random process
- Then we can estimate the (common) expected value from the sample and its presumed spatial structure

Problems with first-order stationarity

- It is often not plausible:
  1. We observe the mean value to be different in several regions (strata)
  2. We observe a regional trend
- In both cases there is a **process** that is not stationary which we can **model**, so ...
- Solution 1: model the strata or trend, then the **residuals** may be first-order stationary (leads to **Kriging with External Drift** or **Regression Kriging**)
- Solution 2: model a varying mean along with the local structure (leads to **Universal Kriging**)
- Solution 3 (Matheron 1965): Study the **differences** between values, not the **values** themselves, and in a **"small"** region

Spatial covariance

- So far we have only estimated a **mean** value, like with a non-regionalized variable. But we want to include the key idea that **nearby observations may be correlated**.
- Just like any other variables, the set of random variables making up the regionalized variable may have **covariance**, i.e. one may be related to another (positively or negatively).
- It's **one realisation per point**, but **each point is a different realisation**, so in some sense they are *different* variables, which then have a covariance.
- **Key Insight:** Under certain assumptions (see below), this covariance can be considered to **depend only on the separation between the points**, and possibly the **direction**.

Commentary

We now introduce the **key insight** in local spatial dependence: **auto-covariance**, i.e. the covariance of a variable **with itself**. How can this be? Patience, and all will be revealed ...

Covariance

- Recall from non-spatial statistics: the **sample covariance** between **two variables** z_1 and z_2 observed at n points is:

$$\hat{C}(z_1, z_2) = \frac{1}{n} \sum_{i=1}^{n} (z_{1_i} - \bar{z}_1) \cdot (z_{2_i} - \bar{z}_2)$$

- **Spatial version**: there is only **one variable** x:

$$\hat{C}(\mathbf{x}_1, \mathbf{x}_2) \quad = \quad E[\{Z(\mathbf{x}_1) - \mu(\mathbf{x}_1)\} \cdot \{Z(\mathbf{x}_2) - \mu(\mathbf{x}_2)\}]$$

- Because of first-order stationarity, the expected values are the same, so:

$$\hat{C}(\mathbf{x}_1, \mathbf{x}_2) \quad = \quad E[\{Z(\mathbf{x}_1) - \mu\} \cdot \{Z(\mathbf{x}_2) - \mu\}]$$

Second-order stationarity (1) – At one point

- Problem: The covariance **at one point** is its **variance**:

$$\sigma^2 = E[\{Z(\mathbf{x}_i) - \mu\}^2]$$

- This can not be estimated from one sample (of the many hypothetical realisations) [we had the same problem with estimating the mean]

- Solution: **assume that the variance at any point is the same finite value at all points**.

- Then we can estimate from the sample the *a priori* variance of the process (i.e. the covariance at a point) by lumping the random variables together.

- This assumption is **second-order stationarity**

Second-order stationarity (2) – Over the spatial field

- **Problem**: The covariance equation as written is between all the points in the field. It is huge! And again, there is no way to estimate these from just one point pair per variable pair.

- **Solution** (the key insight): **Assume that the covariance between points depends only on their separation** (and not on their actual location or individuality).

- Then we can **estimate** their covariance from a **large number of sample pairs**, all **separated** by (approximately) the same separation vector $\mathbf{h}$ (distance, possibly with direction).

Derivation of covariance function

- **Autocovariance** ('auto' = same regionalized variable), at a separation $\mathbf{h}$:

$$\begin{aligned} C[Z(\mathbf{x}), Z(\mathbf{x}+\mathbf{h})] &= E[\{Z(\mathbf{x}) - \mu\} \cdot \{Z(\mathbf{x}+\mathbf{h}) - \mu\}] \\ &= E[\{Z(\mathbf{x})\} \cdot \{Z(\mathbf{x}+\mathbf{h})\} - \mu^2] \\ &\equiv C(\mathbf{h}) \end{aligned}$$

- **Autocorrelation**: Autocovariance normalized by total variance σ^2, which is the covariance at a point:

$$\rho(\mathbf{h}) = C(\mathbf{h}) / C(\mathbf{0})$$

- **Semivariance**: deviation of covariance at some separation from total variance:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

Characteristics of Spatial Correlation functions

- symmetric: $C(\mathbf{h}) = C(-\mathbf{h})$ etc.
- range of $\rho(\mathbf{h}) \in [-1 \cdots 1]$
- Positive (covariance) or negative (variogram) semi-definite matrices; this restricts the choice of models
- Continuity, especially at $\mathbf{0}$. But this is often not observed: the "nugget" effect; we will examine this in variogram analysis

Problems with second-order stationarity

- It is often not plausible; in particular the covariance often increases without bound as the area increases.
- Solution 1 (Matheron 1965): Study the **differences** between values, not the **values** themselves, and in a **"small"** region; then the covariances may be bounded
- Solution 2: model **semi-variance**, not **co-variance** (see below)

The Intrinsic Hypothesis

- Replace mean **values** with mean **differences**, which are the same over the whole field, at least within some 'small' separation $\mathbf{h}$. Then the expected value is 0:

$$E[Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})] = 0$$

- Replace **covariance of values** with **variances of differences**:

$$\mathrm{Var}[Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})] = E[\{Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\}^2] = 2\gamma(\mathbf{h})$$

- The equations only involve the **difference in values** at a separation, not the **values**; so the changing mean and increasing variance over the area are compensated.

- This is the **intrinsic hypothesis**

Using the experimental variogram to model the random process

- Notice that the semivariance of the separation vector $\gamma(\mathbf{h})$ is now given as the estimate of covariance in the spatial field.
- So it models the **spatially-correlated component** of the **regionalized variable**
- We must go from the **experimental variogram** to a **variogram model** in order to be able to model the random process at any separation.

Topic 3: Models of spatial covariance

1. Recall (lecture 2): The **empirical variogram** is often computed as:

$$\overline{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} [z(\mathbf{x}_i) - z(\mathbf{x} + \mathbf{h}_i)]^2$$

   - $m(\mathbf{h})$ is the number of **point pairs** separated by vector $\mathbf{h}$, in practice some range
   - Point-pairs indexed by i; the notation $\mathbf{x} + \mathbf{h}_i$ means the "tail" of point-pair i, i.e. separated from the "head" $\mathbf{x}_i$ by the separation vector $\mathbf{h}$.

2. In the previous section, we derived the **theoretical variogram** as a **continuous function** of semi-variance vs. separation $\mathbf{h}$:

$$\gamma(\mathbf{h}) = 0.5 \cdot E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2]$$

How can we relate these?

From experimental to theoretical variogram

If we **assume** that there is a **spatially-autocorrelated process** which depends on a theoretical variogram, we can then try to **estimate its form and parameters** from the **empirical variogram**.

But first we need to develop some **models of spatial covariance**, in other words, what forms can the covariance function take?

Commentary

We can imagine many functions of semivariance based on separation; but not all of these are:

- derived from plausible spatial processes;
- mathematically valid for interpolation ("kriging")

The set of functions that are useful are called **authorized** models.

We will first see what properties they must have, then some examples, and finally visualize them.

Authorized Models

- Any variogram function must be able to model the following:
  1. **Monotonically increasing**
     - possibly with a single fluctuation (**hole**) ...
     - ...or a **periodic** fluctuation
  2. Constant or asymptotic maximum (**sill**)
  3. Non-negative intercept (**nugget**)

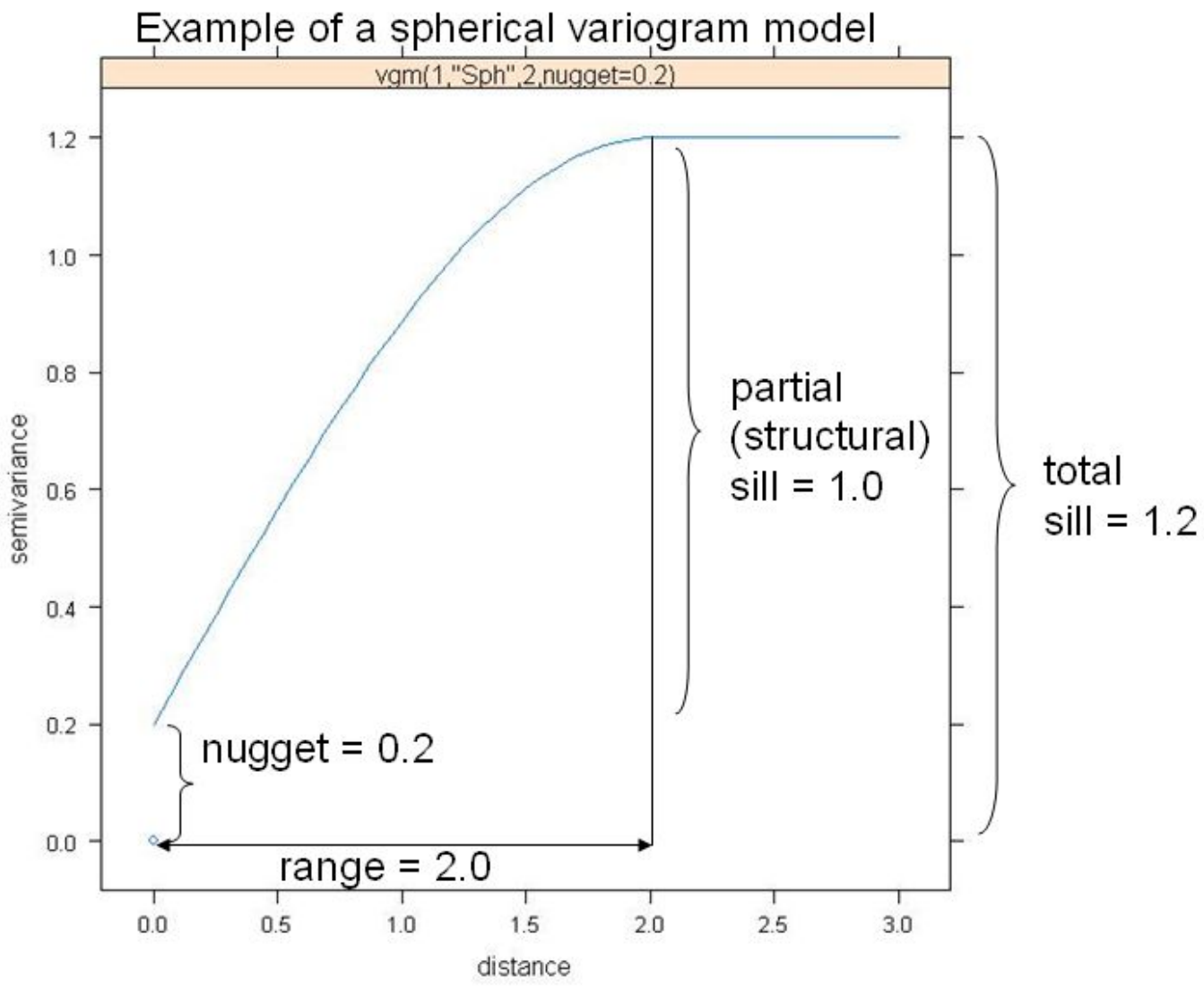
- Variograms must obey mathematical constraints so that the resulting kriging equations are solvable (e.g., **positive definite** between-sample covariance matrices).

- The permitted functions are called **authorized models**.

Commentary

Before we look at the various authorized model forms, we show one of the simplest models, namely the **spherical** model, with the various **model parameters** labelled.

A variogram model, with parameters



Derivation of authorized models

- Assume the **underlying stochastic process**
- Derive the **covariance function** that results from values assigned to points according to this process.
  - E.g. spherical model is derived from the intersection of random spheres of a given size: how much will they overlap?
- Convert to **variogram function**

These variogram functions are generally available in geostatistical modelling and interpolation computer programs.

Examples of authorized models (1/4)

- Group 0: **Pure nugget** (no spatial structure): $\gamma(h) = c_0, \forall h > 0$
  - interpolation is impossible, so ...
  - ... **the sample mean estimates every point**

- Group 1: **Unbounded** models: variance increases with area; no sill or range
  - Suggests that **study area is smaller than the range of spatial dependence** of the process
  - **Power** model $\gamma(h) = wh^{\alpha}, 0 < \alpha < 2$
  - α controls shape, i.e. the rate of change of the variance with distance
    1. $\alpha = 1$ linear: same
    2. $\alpha < 1$ convex: increasing
    3. $\alpha > 1$ concave: decreasing

Examples of authorized models (2/4)

- Group 2a: **Bounded** models: variance *reaches* a **sill** at some **range**

1. **Bounded linear**, sill c, range a; **only authorized for 1D** (e.g. transects)

$$\gamma(h) = \begin{cases} c(h/a) & : \quad h < a \\ c & : \quad h \geq a \end{cases}$$

2. **Spherical**, sill c, range a

$$\gamma(h) = \begin{cases} c\left\{\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\} & : \quad h < a \\ c & : \quad h \geq a \end{cases}$$

These two models are quite simular to the common **spherical** models (see previous slide); they differ mainly in the shape of the "shoulder" transition to the sill:

3. **Circular**, sill c, range a

$$\gamma(h) = \begin{cases} c \left\{ \frac{2h}{\pi a} \sqrt{1 - \left(\frac{h}{a}\right)^2} + \frac{2}{\pi} \arcsin\left(\frac{h}{a}\right) \right\} & : \quad h < a \\ c & : \quad h \geq a \end{cases}$$

4. **Pentaspherical**, sill c, range a

$$\gamma(h) = \begin{cases} c \left\{ \frac{15}{8}\frac{h}{a} - \frac{5}{4}\left(\frac{h}{a}\right)^3 + \frac{3}{8}\left(\frac{h}{a}\right)^5 \right\} & : \quad h < a \\ c & : \quad h \geq a \end{cases}$$

Examples of authorized models (3/4)

- Group 2b: **Bounded asymptotic** models: variance **approaches** a **sill** at some **effective range** (by convention, where $\gamma = 0.95c$)

4. **Exponential**, sill c, **effective range** $3a$

$$\gamma(h) = c\{1 - e^{(-\frac{h}{a})}\}$$

E.g. if the effective range is estimated as 120, the parameter a is 40.
Note: Some authors (e.g. Goovaerts) use $3h$ instead of h in this equation, in which case a is the effective range itself.

Examples of authorized models (4/4)

5. **Gaussian**, parameters as for exponential, but **effective range** is $\sqrt{3}a$:

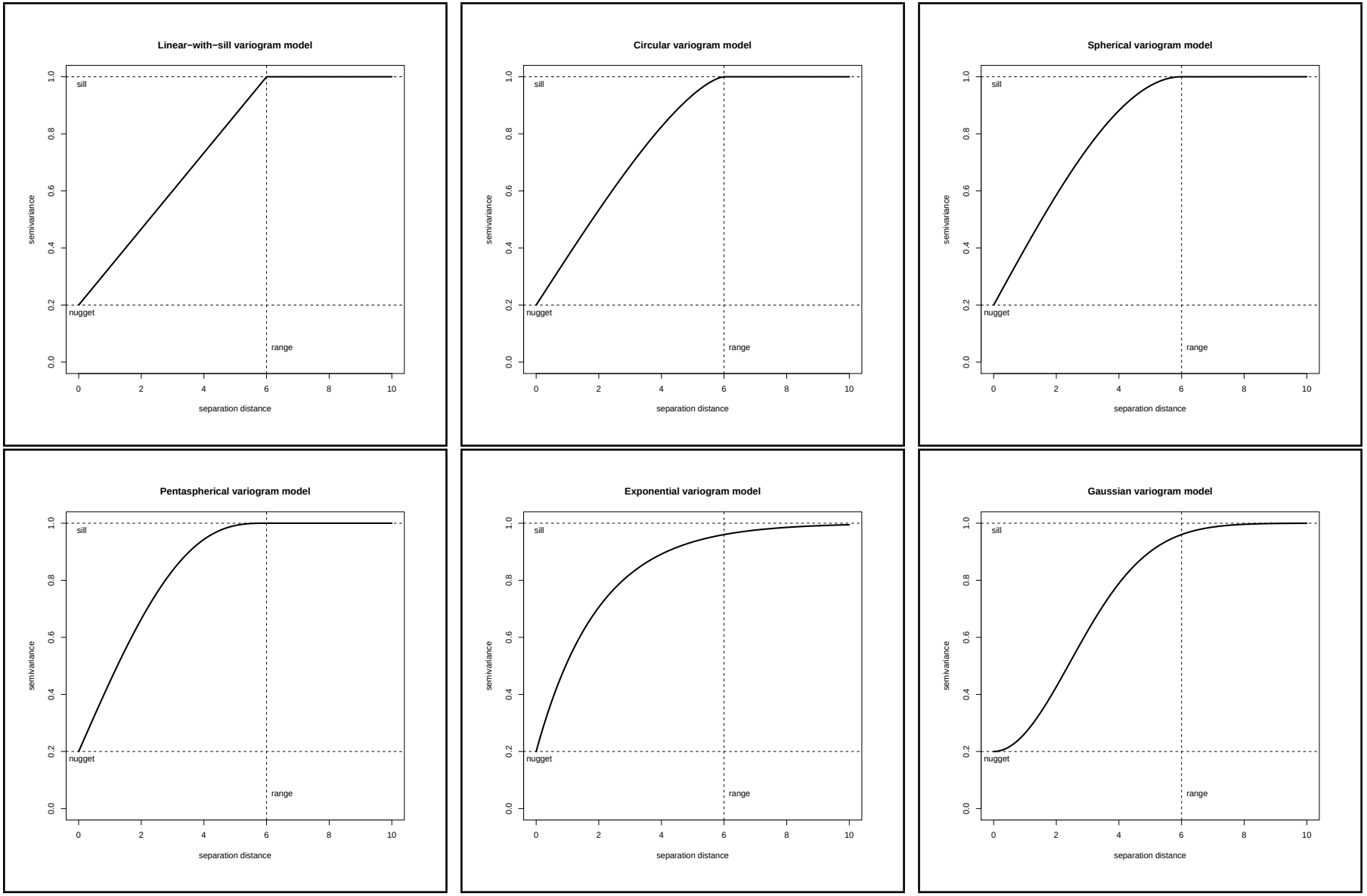
$$\gamma(h) = c\{1 - e^{-\left(\frac{h}{a}\right)^2}\}$$

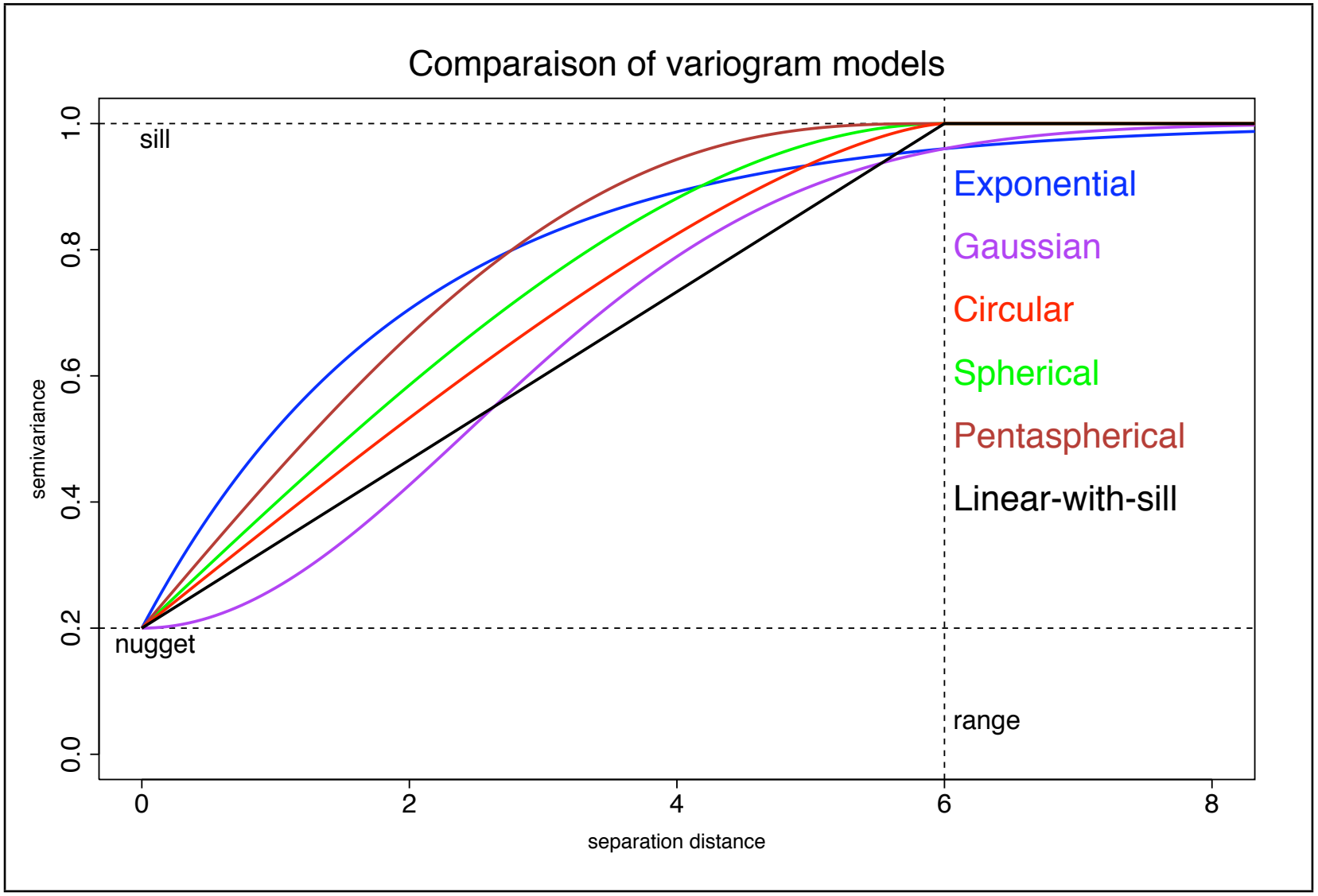
E.g. if effective range is estimated as 120, the parameter a is $120/\sqrt{3} \approx 69$

Note: Some authors (e.g. Goovaerts) use $3h$ instead of h in this equation, in which case a is the effective range itself, as in the exponential model

This model differs from the exponential model at small separations; semi-variances are close to zero in a "halo" around each point.

Graphs of authorized variogram models

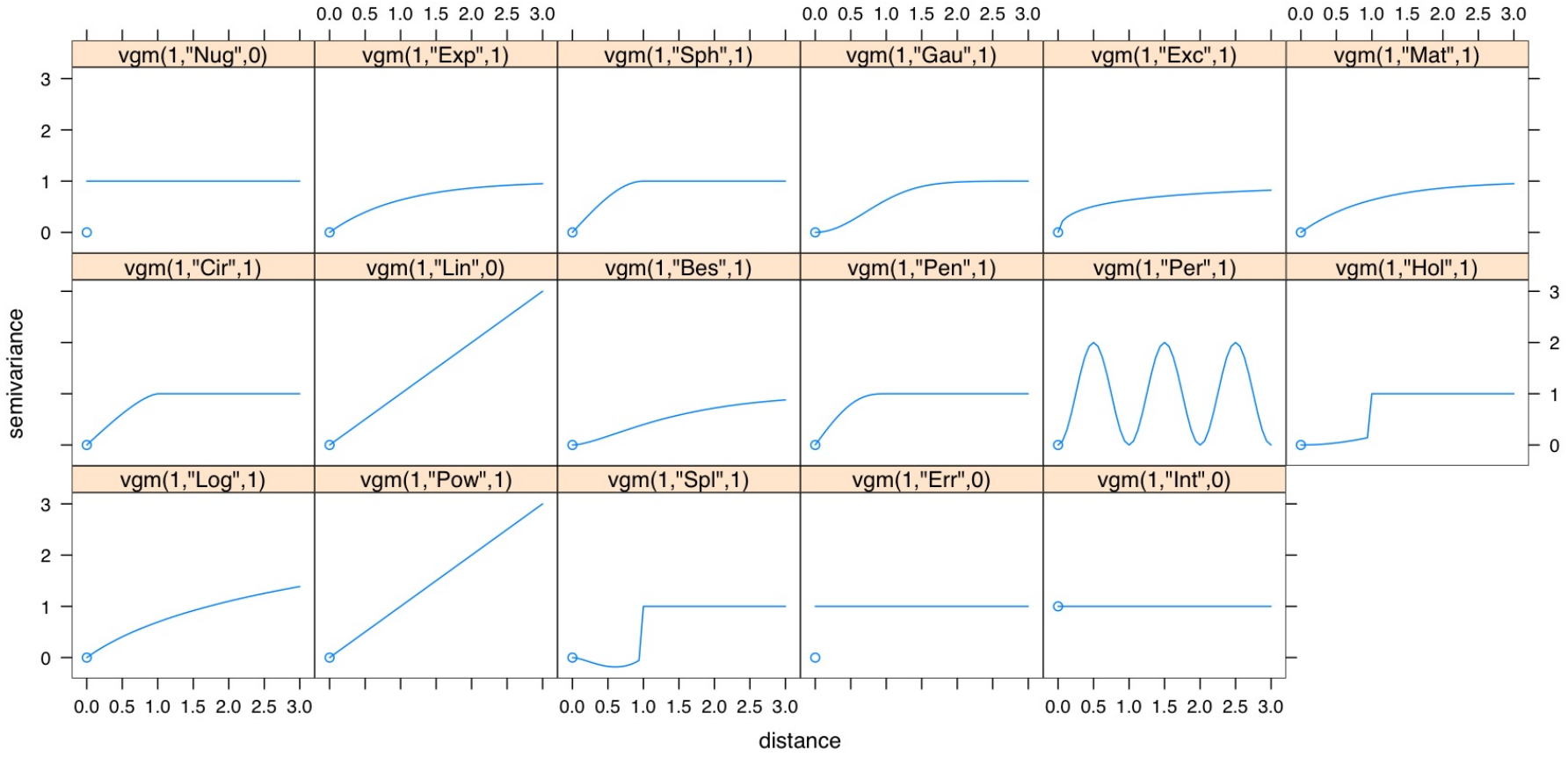




Models vary considerably, from origin to range

Note effective range (95% of asymptotic sill) for exponential & Gaussian models

Models available in gstat



Model families

The concept of a variogram model can be **generalized** into **model families** require an additional parameter to adjust their shape.

These are able to adjust their **smoothness** to correspond to the nature of the process that generated the random field.

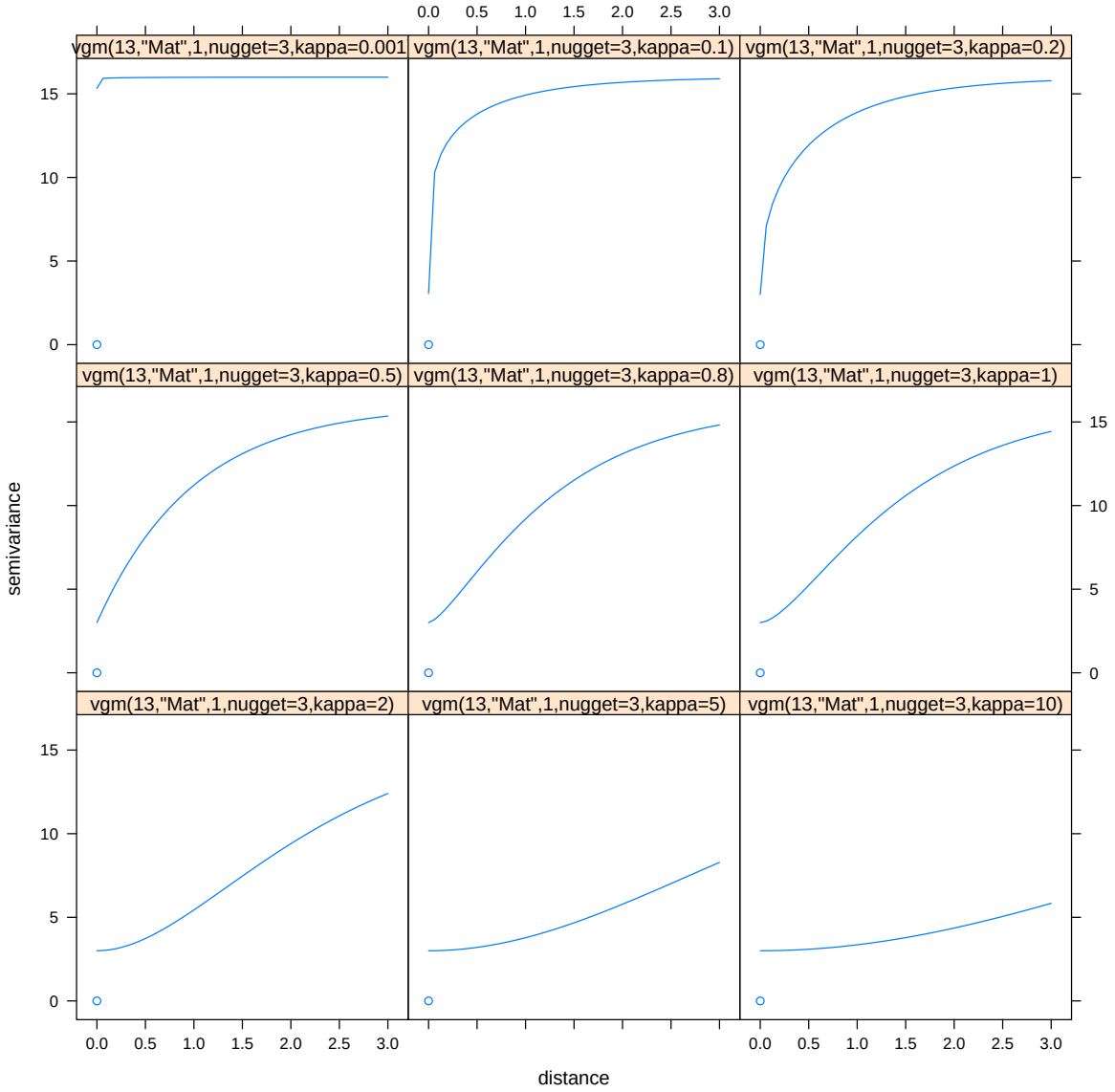
Sophisticated automatic fitting methods can adjust the smoothness, as well as the sill and range, so they are applicable in "black box" variogram fitting.

An example is the Matérn model, which generalizes the Exponential, Power, Logarithmic and Gaussian models

Reference:

Minasny, B., & McBratney, A. B. (2005). *The Matérn function as a general model for soil variograms*. Geoderma, 128, 192-207.

Matérn model class, with varying smoothness parameter



Topic 4: Variogram analysis and model fitting

From the **empirical variogram** we now derive a **variogram model** which expresses semivariance as a **function of separation vector**.

The model allows us to:

- Infer the characteristics of the **underlying process** from the functional form and its parameters;
- Compute the semi-variance between *any* point-pair, separated by *any* vector ...
- ...which is used in an **'optimal' interpolator** ("kriging") to predict at unsampled locations.

Modelling from one realisation of a random field

If we accept the theory of **random functions** which generated **spatial fields** we suppose there's a **true variogram** (covariance function) of the **spatially-correlated random process** ...

- ... of which we have only one **realisation** in nature ...
- ... and (in addition) we have only a **sample** of that realisation.

So, we try to **estimate** it from the **experimental variogram**.

Commentary

There are two steps to modelling spatial dependence from the experimental variogram:

1. Decide on the **variogram model** form, e.g. spherical, exponential, Gaussian ...
2. ... **Parameterize** the chosen model, i.e. **fit** it to the experimental variogram.

Recall: there is no **objective** way to do either of these! because we have only **one realization** of the assumed **random field**.

This is then a controversial and difficult step.

Please rememember: **your** data set is only **one of many** from similar studies. So don't put too much emphasis on the one dataset; compare it with what you know about the **process**.

Choosing a model (1)

The empirical variogram should be **one realization of a random process**.

So, the first question to ask when deciding on a model form is: **what do we expect from the process** that is supposed to be responsible for the spatial structure represented in the variogram?

Each model has some defining characteristics:

- **Exponential**: First-order autoregressive process: values are random but with dependency on the nearest neighbour; boundaries according to a Poisson process
- **Gaussian**: as exponential, but with strong close-range dependency, very smooth at each point.

(continued . . .)

Choosing a model (1) – continued

- **Spherical**, **circular**, **pentaspherical**: **Patches** of similar values; patches have similar size $\approx$ range) with **transition** zones (overlap of processes);

  These three model forms differ mainly in the "shoulder" transition to the sill.

- **Bounded linear**: Autocorrelation decreases linearly with distance, to zero at some range

- **Unbounded**, e.g. **power model**: range of process larger than our sample area

- **Periodic**: as it says

Examples of simulated fields

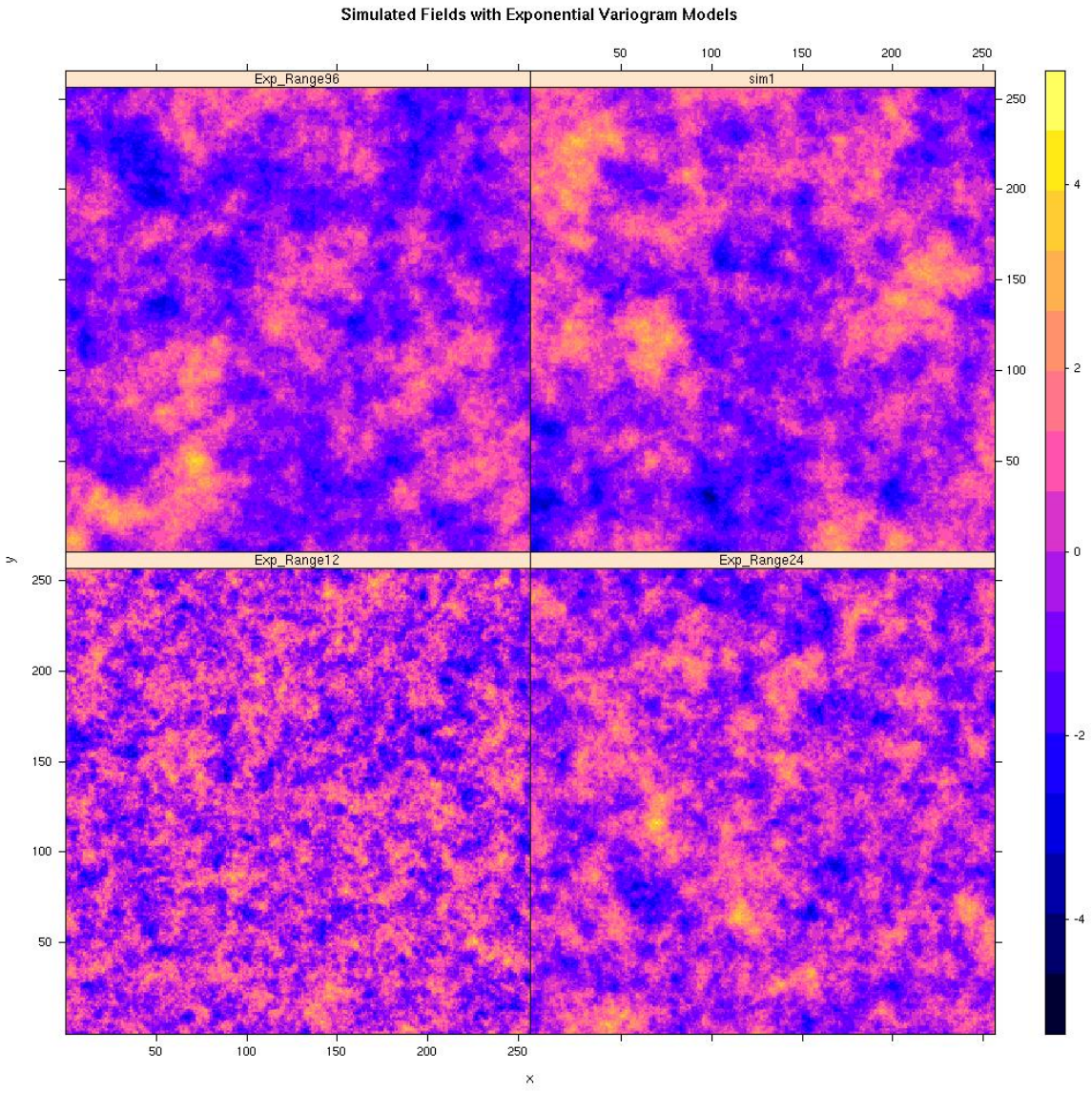
The following graphs show **simulations** of a random field, assuming the underlying process is according to the two models.

1. **Exponential**, from short– to long–range dependence

2. **Spherical**, long–, medium–, short–range; also compare with **pure nugget**

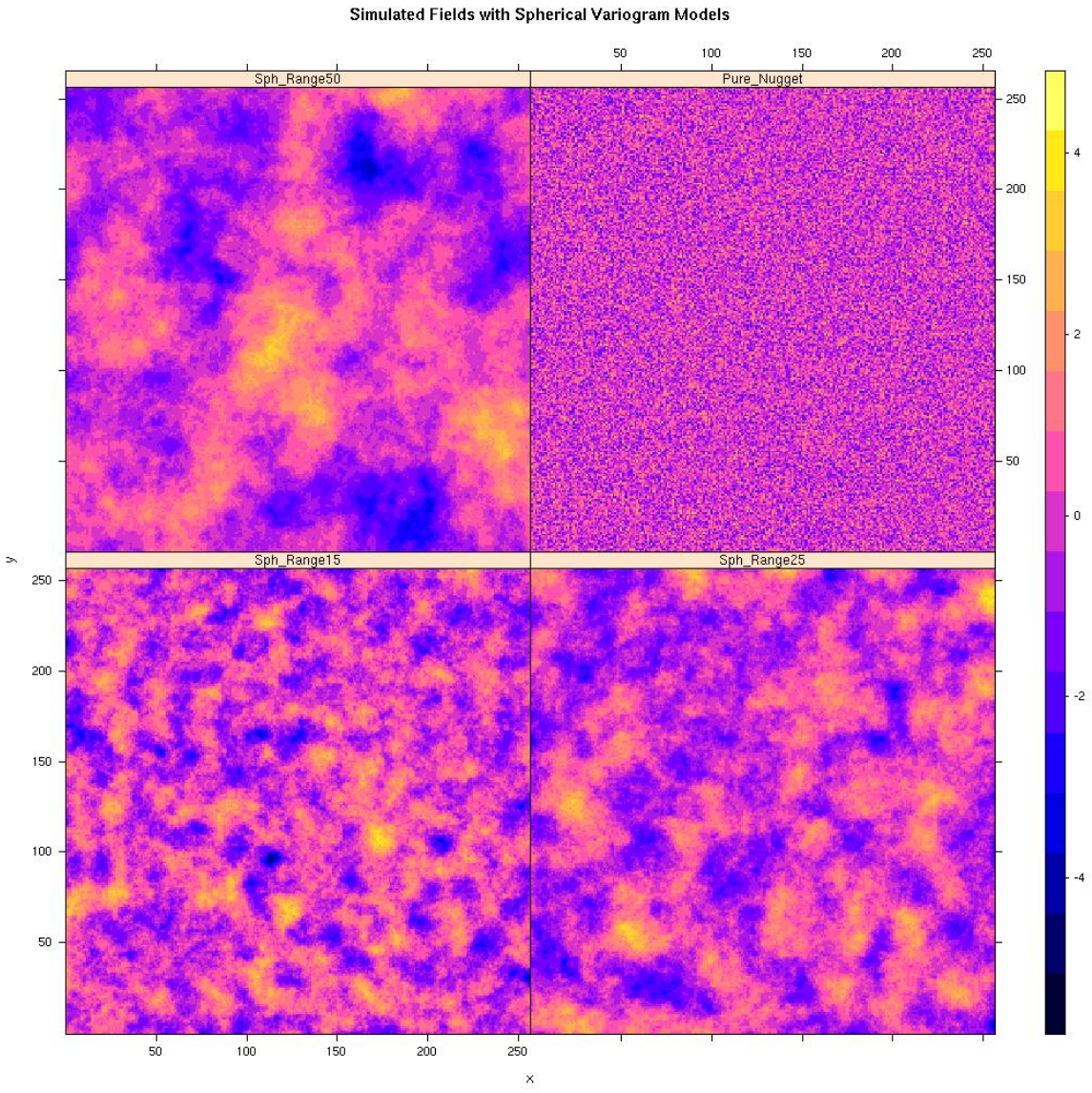
(Simulations computed with `krige` method of `gstat` package in `R`.)

Simulations of regionalised variables: Exponential



Effective Ranges 96, 48, 24, 12 on 256x256 grid

Simulations of regionalised variables: Spherical



Ranges 50, 0 (pure nugget), 15, 25 on 256x256 grid

To check your understanding . . .

Q11 : *Which model form shows better-defined* **patches**, *the exponential or spherical?* *Jump to A11* •

Q12 : *What happens to the patch size as the* **range parameter** *increases?* *Jump to A12* •

Choosing a model (2)

- **Which has been successfully applied with this kind of data?**
  (This is evidence for the nature of this kind of process)

- **What do we expect from the supposed process?** if we have some other evidence of its spatial behaviour.
  For example, a Gaussian model might be expected for a phenomenon which physically must be **continuous** at each point, e.g. the surface of a ground-water table.

- **Visual estimate** of functional form from the variogram

Choosing a model (3)

- **Fit various models**, pick the **statistically-best fit**
  - This **ignores all prior information** about the process
  - Depends heavily on the experimental variogram parameters, e.g. cutoff, bin width
  - Depends also on the **weighting function** for the typical **weighted least-squares** non-linear fitting method.

Conclusion: This is **not** by itself a good method to choose a model.

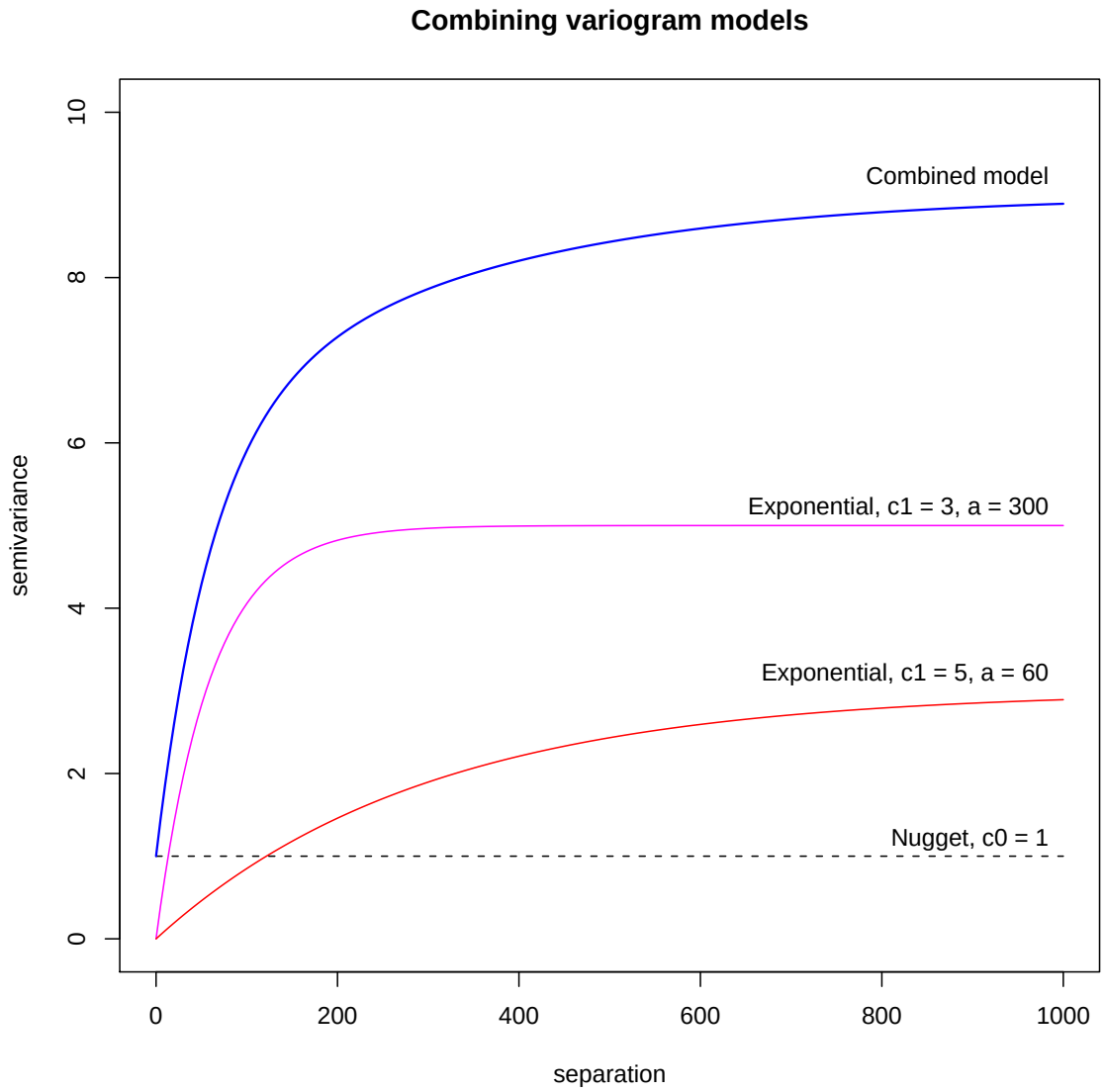
Combining models

- Any **linear combination** of authorized models is also authorized
- Models > 1 spatial structure at different distances
- Common example: **nugget + structural**
- e.g. nugget + exponential

$$\gamma(h) = c_0 + c_1\{1 - e^{(-\frac{h}{r})}\}$$

- Structure at two ranges: e.g. nugget + exponential + exponential
- Prefer a simpler (more **parsimonious**) model, but sometimes the structure calls for a combination.

Combining variogram models



Evidence of both short range $(3a = 180)$ and long range $(3a = 900)$ phenomena

Fitting the model

Once a **model form** is selected, then the **model parameters** must be adjusted for a 'best' fit of the experimental variogram.

- **By eye**, adjusting parameters for good-looking fit
  - Hard to judge the relative value of each point
- **Automatically**, looking for the best fit according to some objective criterion
- In both cases, favour sections of the variogram with **more pairs** and at **shorter ranges** (because it is a **local** interpolator).
- **Mixed**: adust by eye, evaluate statistically; or vice versa

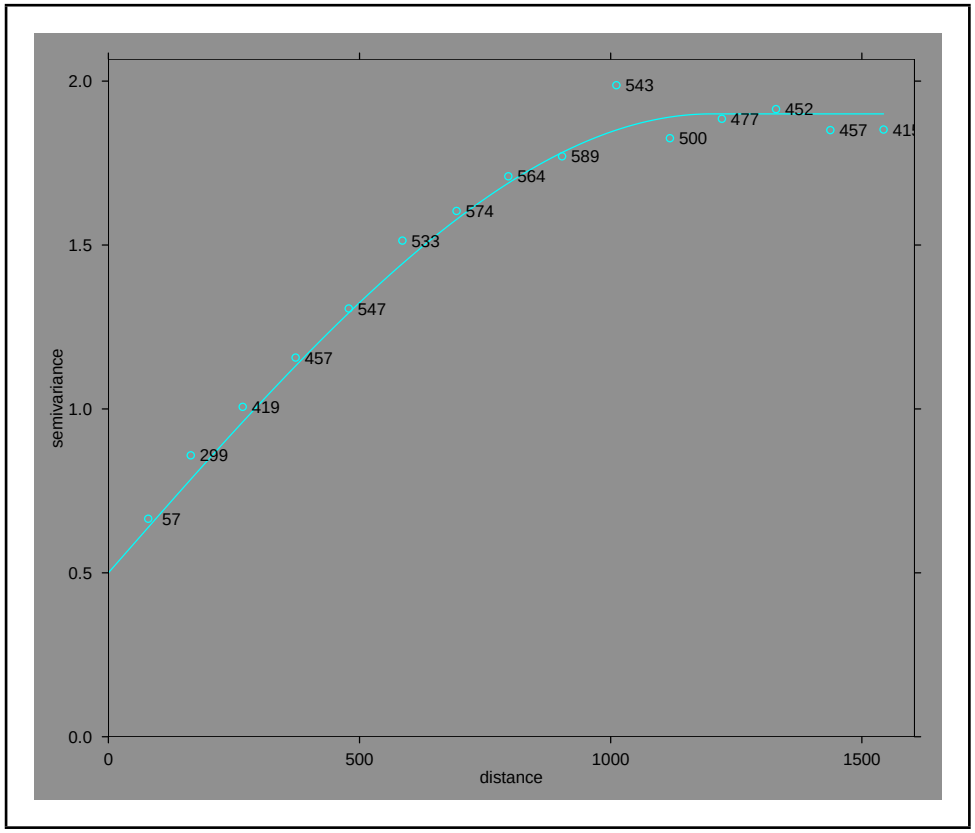
Commentary

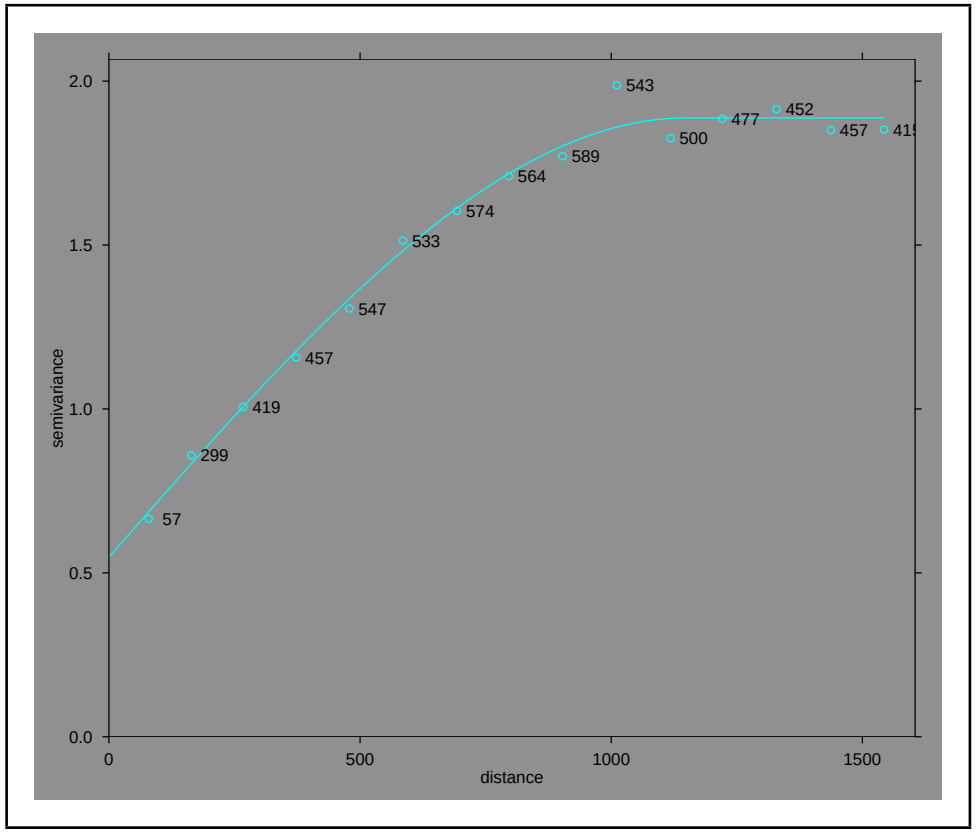
We now look at two models that were fitted first by eye and then automatically.

In the first case there is little dispute; the second is much more difficult to estimate and fit.

The next slide shows the "easy" case; after a discussion of automatic fit options the "hard" case is shown.

Meuse dataset, Log(Cd): Eyeball (L) vs. automatic (R) fits





By eye: $c_0 = 0.5, c_1 = 1.4, a = 1200$; total sill $c_0 + c_1 = 1.9$

Automatic: $c_0 = 0.548, c_1 = 1.340, a = 1149$; total sill $c_0 + c_1 = 1.888$

The total sill was almost unchanged; `gstat` raised the nugget and lowered the partial sill of the spherical model a bit; the range was shortened by 51 m.

Statistical measures of variogram fit

- Minimize the deviation of the variogram model from the sample variogram, **weighted** by the **number of point pairs** (more → more important) and **separation** (closer → more important)

$$\sum_{j=1}^{b} \frac{N_j}{h_j^2} [\hat{\gamma}(\mathbf{h}_j) - \gamma(\mathbf{h}_j)]^2$$

- The $\hat{\gamma}()$ are the **experimental** semivariances for a distance class
- The $\gamma()$ are the **modelled** semivariances for a distance class
- b is the number of bins
- The weighting factor is not theoretical, but has proved good in practice
- Other weightings are possible.

Commentary

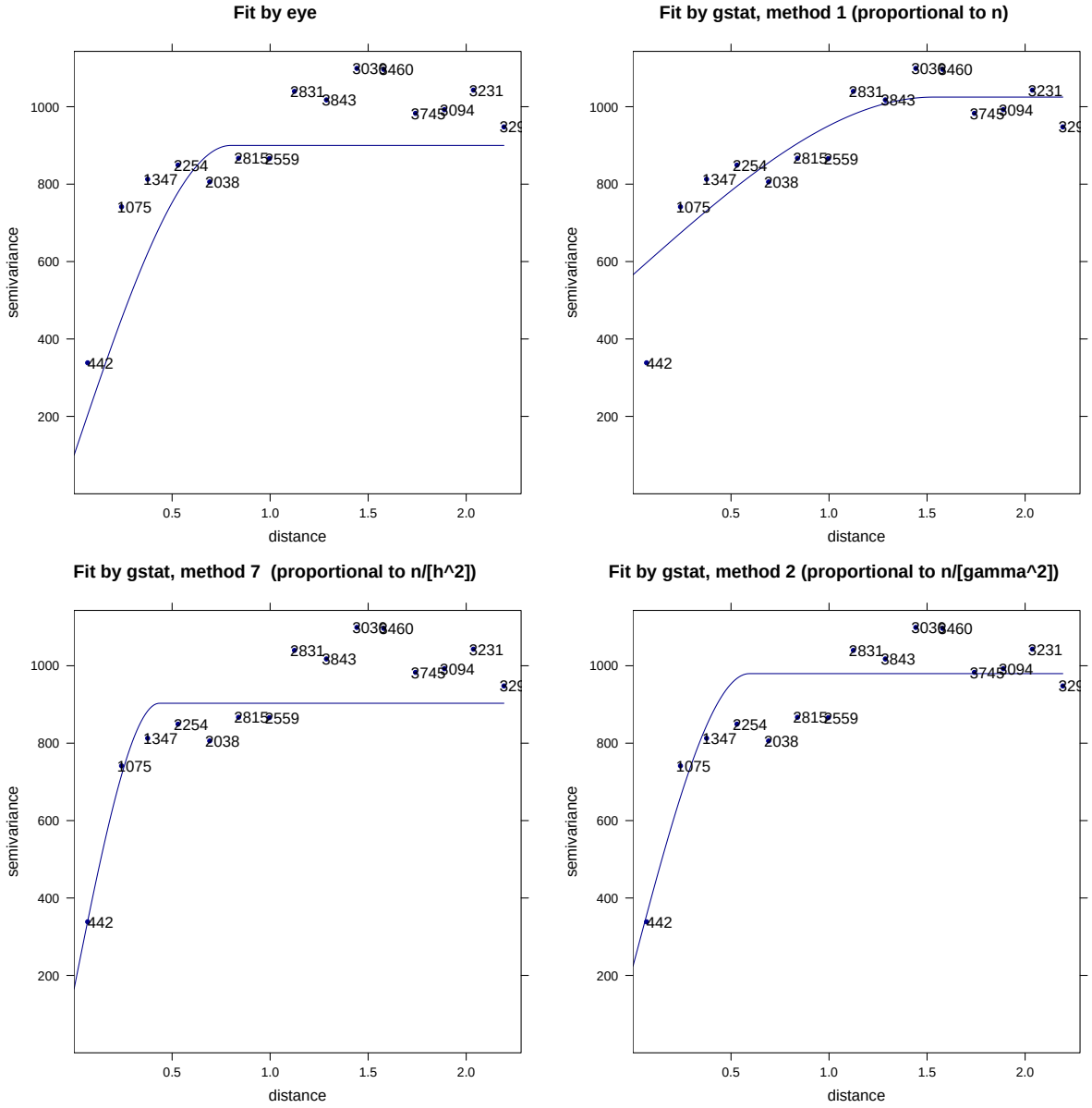
"Automatic" fitting should always be guided by the analyst. If the variogram is quite consistent across the range, almost any automatic fit will be the same. Otherwise, different fitting criteria will give different results.

The next page shows three different automatic fits:

1. Weights only proportional to **number** of point-pairs in variogram bin (`gstat` method 1)
2. Weights proportional to **number** of point-pairs in variogram bin and inversely proportional to the square of the **separation** (`gstat` method 7, the default);
3. Weights proportional to **number** of point-pairs in variogram bin and inversely proportional to the square of the **semivariance** (`gstat` method 2);

All of these weighting schemes have been used; there is no theoretical basis for choosing one of them.

Jura Zn content; eyeball and three automatic fitting methods



What sample size to fit a variogram model?

- Can't use non-spatial formulas for sample size, because spatial samples are correlated, and each sample is used multiple times in the variogram estimate
- No way to estimate the actual error in the variogram fit, since we have only one realisation of the random field
- Stochastic simulation from an assumed random field with a known variogram 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.

This is **very worrying for many environmental datasets** (soil cores, vegetation plots, ...) especially from short-term fieldwork, where sample sizes of 40 – 60 are typical. Should variograms even be attempted on such small samples?

Exercise

At this point you should complete **Exercise 3: Modelling spatial structure** which is provided on the module CD.

This should take several hours.

1. Trend surfaces
2. Theory of random fields
3. Models of spatial covariance
4. Variogram analysis; variogram model fitting

In all of these 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.

Answers

Q1 : *Describe a process in your field of study that would be expected to show this behaviour, i.e. the attribute value depends on relative geographic position within a spatial field* •

A1 : *(No single answer)* *Return to Q1* •

Answers

Q2 : *What is the change in clay content for each km towards the North?* •

A2 : -0.45%; *see the coefficient for* `UTM_N` *in the model summary.* *Return to Q2* •

Q3 : *What proportion of the residuals are between* -5.106 *and* $+3.607$*?* •

A3 : *Half of them; these are the* 1st *and* 3rd *quartiles, respectively.* *Return to Q3* •

Answers

Q4 : *How well do the fitted values match the observed values?* •

A4 : *Very poorly; in particular the fitted values are clustered around 30, 45 and 55% clay (vertical "clouds"), which pattern is not seen in the observed values.* *Return to Q4* •

Answers

Q5 : *Is there any pattern of the residuals vs. fitted values?* •

A5 : *Yes, they are more variable near fitted = 55%; there are some very poorly-modelled points in that range.* *Return to Q5* •

Q6 : *Are the residuals normally-distributed?* •

A6 : *No, the upper tail is too short and the lower tail too long.* *Return to Q6* •

Q7 : *Which observations are marked by the diagnostic procedure as being outliers in this linear model fit?* •

A7 : *‘ 84 amd 145 (very large negative residuals, i.e. actual much less than fitted); 2 (very large positive residual, actual much higher than fitted)* *Return to Q7* •

Answers

Q8 : *How much improvement is there with the second-order, compared to the first-order, trend surface? Comment on the adjusted R^2, the fitted vs. observed values, and the diagnostic plots of the residuals.* •

A8 : *There is only a 2% improvement in adjusted R^2 (from 0.499 to 0.519); the fitted vs. observed values and the diagnostic plots of the residuals are very similar in both cases. The same outliers are flagged in the residual plots of both orders.* *Return to Q8* •

Answers

Q9 : *How similar are the* **attribute values** *at any* **specific location***?* •

A9 : *The attribute values are in general not the same over the four realizations at any given location.* *Return to Q9* •

Q10 : *How similar are the* **patterns***? Compare the patch sizes, sharpness of transitions between patches, etc.* •

A10 : *However, the pattern is very similar; only the specific locations of high and low patches is different. These look like they could be four squares cut out of some larger area.* *Return to Q10* •

Answers

Q11 : *Which model form shows better-defined* **patches**, *the exponential or spherical?* •

A11 : *The spherical model has a patchier appearance, with sharper transition zones. The exponential has clear "hot spots" but the transition is more diffuse.* *Return to Q11* •

Q12 : *What happens to the patch size as the* **range parameter** *increases?* •

A12 : *The patch size increases accordingly.* *Return to Q12* •