Spatial Data Analysis in R

Spatial Dependence and Autocorrelation 1

Eco 697DR – University of Massachusetts, Amherst – Spring 2022
Michael France Nelson

Deck 1 concepts

Spatial dependence overview

- Causes
- Implications

Quantifying spatial dependence: spatial statistics and geostatistics

- Moran's I and correlograms
- Semivariance and variograms

Geostatistics and spatial statistics

Inference paradigms: geostatistics and spatial statistics

- prediction and inference
- Variance and correlation

Common goals

- Estimate spatial correlation, variance, and covariance.
- Use statistical techniques to make informed guesses.
- Identify characteristic scales.

Geostatistics

Geostatistics is a branch of spatial statistics with an origin in mining applications.

JOURNAL OF THE CHEMICAL METALLURGICAL & MINING SOCIETY OF SOUTH AFRICA

The Society, as a body, is not responsible for the statements and opinions advanced in any of its Publications

Vol. 52 No. 6 DECEMBER 1951 Price 6/-

A STATISTICAL APPROACH TO SOME BASIC MINE VALUATION PROBLEMS ON THE WITWATERSRAND

By D. G. KRIGE, M.Sc. (Eng.) (Rand)

Causes of dependence

Tobler's 1st law.

Why might nearby things be related? The F +F book identifies 2 types of dependence sources:

- "Endogenous mechanisms are those that directly occur from the organism or processes being considered, which result in spatial pattern."
- "Exogeneous mechanisms, in contrast, are those that occur outside of the organism or process being measured, such as spatial aggregation of resources or environmental gradients used by the organism of interest, which is sometimes referred to as "indirect" mechanisms…"

Why do we care about spatial dependence?

Statistical considerations:

- Independence
- Sample size
- Degrees of freedom

Phenomenological considerations

- Dependence may be focus of interest.
- Why might you care about dependence in your own research?

Statistical Reasons to Care

Non-independence of observations

- Non-independent observations contain less information
- Pseudoreplication: non-independence causes the effective degrees of freedom to be smaller than n – 1.

Dependence mostly affects measures of spread and confidence/significance (not center).

Example: sample standard deviation and standard error of the mean

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Quantifying Autocorrelation: Moran's I and Correlograms

Quantification: Moran's I

A statistic related to variance, covariance, and correlation:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

The formula is a lot to digest, but we've seen all the components before.

We'll work through it to gain some insight.

Dissecting Moran's I: variance and covariance

Approximate verbal definitions:

Variance answers the question: "How much do the values in a collection of numbers differ from their mean, *on average*?"

$$\text{variance} = \frac{\text{sum of squared deviations from the mean}}{N}$$

Covariance answers the question: "How well-coordinated are variations in two variables: x and y?"

Variance and Covariance Forumlae

$$var(x) = \frac{\sum_i (x_i - \bar{x})^2}{N}$$

$$cov(x_1, x_2) = \frac{\sum_i (x_{1i} - \bar{x_1})\,(x_{2i} - \bar{x_2})}{N}$$

$$cov(x_1, x_1) = var(x_1)$$

Things to note:

- They both have N in the denominator: they are averages.
- They both contain sums of squared deviations.
- Can they be negative?

Pearson's correlation

A *normalized* version of covariance. Normalized by what?

$$corr(x_1, x_2) = \frac{cov(x_1, x_2)}{\sqrt{var(x_1)var(x_2)}}$$

This resembles a ratio of two covariances…

- Remember that $cov(x_1, x_1) = var(x_1)$.
- If x_1 and x_2 are the same (or if their variability is perfectly coordinated) the numerator and denominator will be equal.
- What if they are perfectly anti-coordinated?
  - Variance is always positive; we never have to deal with complex numbers!

Weight matrix

$$I = \frac{n}{W} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

The Moran's I formula contains *weight matrix terms*, $w_{i,j}$,and the term W.

W is the count of all the elements considered by the sum terms.

- This is a little bit like the N term.

In a sense the W, N, and terms cancel, leaving something that resembles Pearson's correlation coefficient.

Dissecting Moran's I

- We can walk through the components of the formula to see how it's like a correlation coefficient.
- First, let's review variance and covariance essentials

Dissecting Moran's I: Variance

Variance

$$\text{variance} = \frac{\text{sum of squared deviations from the mean}}{N}$$

Sample variance

$$var(x) = \frac{\sum_i (x_i - \bar{x})^2}{N}$$

Dissecting Moran's I: Covariance

Sample covariance

$$cov(x_1, x_2) = \frac{\sum_i (x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{N}$$

Covariance and variance relationship

$$cov(x_1, x_1) = var(x_1)$$

Dissecting Moran's I: Pearson's Correlation

- Pearson's correlation is a normalized version of covariance.
- Why normalize?
  - Covariance can range from negative to positive infinity…
  - Covariances calculated from different sets of variables are not comparable
  - Covariance is in weird units.
  - Normalizing constrains the range: -1 to 1
  - Normalizing makes correlation unitless
- Normalized by What?
  - Variances of the two variables (actually the square root of their product).

Dissecting Moran's I: Correlation

2.4 Pearson's correlation

A *normalized* version of covariance.

$$corr(x_1, x_2) = \frac{cov(x_1, x_2)}{\sqrt{var(x_1)var(x_2)}}$$

Dissecting Moran's I: Complete Formula

3.1 Moran's I

A common formulation for Moran's I is:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Dissecting Moran's I: Variance Part

3.2.1 Variance part

This resembles the reciprocal of a variance formula:

$$\frac{1}{var(x)} = \frac{N}{\sum_i (x_i - \bar{x})^2}$$

Dissecting Moran's I: Variance Part

It's especially easy to see the connection to variance when we invert:

$$var(x) = \frac{\sum_i (x_i - \bar{x})^2}{N}$$

Dissecting Moran's I: Covariance Part

3.2.2 Covariance part

This part looks a little like a covariance formula:

$$\frac{\sum_i \sum_j w_{ij} \left(x_i - \bar{x}\right) \left(x_j - \bar{x}\right)}{W}$$

Note the double summation, the w_i, and the division by W.

Dissecting Moran's I: Weight Matrix Part

3.2.3 Weight matrix part

The w_{ij} part in the summation term:

$$\sum_i \sum_j w_{ij}$$

Is analogous to an **indicator** function for pairs of points within the row-normalized distance matrix of all points *x*: .

Dissecting Moran's I: Weight Matrix Part

Combining the summation term with is analogous to counting the number of pairs of points:

$$\frac{\sum_i \sum_j w_{ij}}{W} = \frac{1}{N_{pairs}}$$

Dissecting Moran's I: Covariance and Weight

It is easier to see the connection to covariance when we present the formula as:

$$\frac{\sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x})}{N_{pairs}}$$

Sample covariance

$$cov(x_1, x_2) = \frac{\sum_i (x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{N}$$

Dissecting Moran' I: Putting it Back Together

3.2.4 Reassembled Moran's I

General idea:

$$I = \left(\frac{N_1}{\text{variance}} \right) \left(\frac{\text{covariance}}{N_2} \right) = \frac{\text{Covariance part}}{\text{Variance part}}$$

In equation form:

$$I = \left(\frac{N_{all}}{\sum_i (x_i - \bar{x})^2} \right) \left(\frac{\sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x})}{N_{pairs}} \right)$$

Moran's I

Moran's I requires a neighborhood definition: its original form is a **global statistic**.

- Answers the question of whether autocorrelation is present.
- Doesn't directly tell us anything about distance

Moran's I at different distances

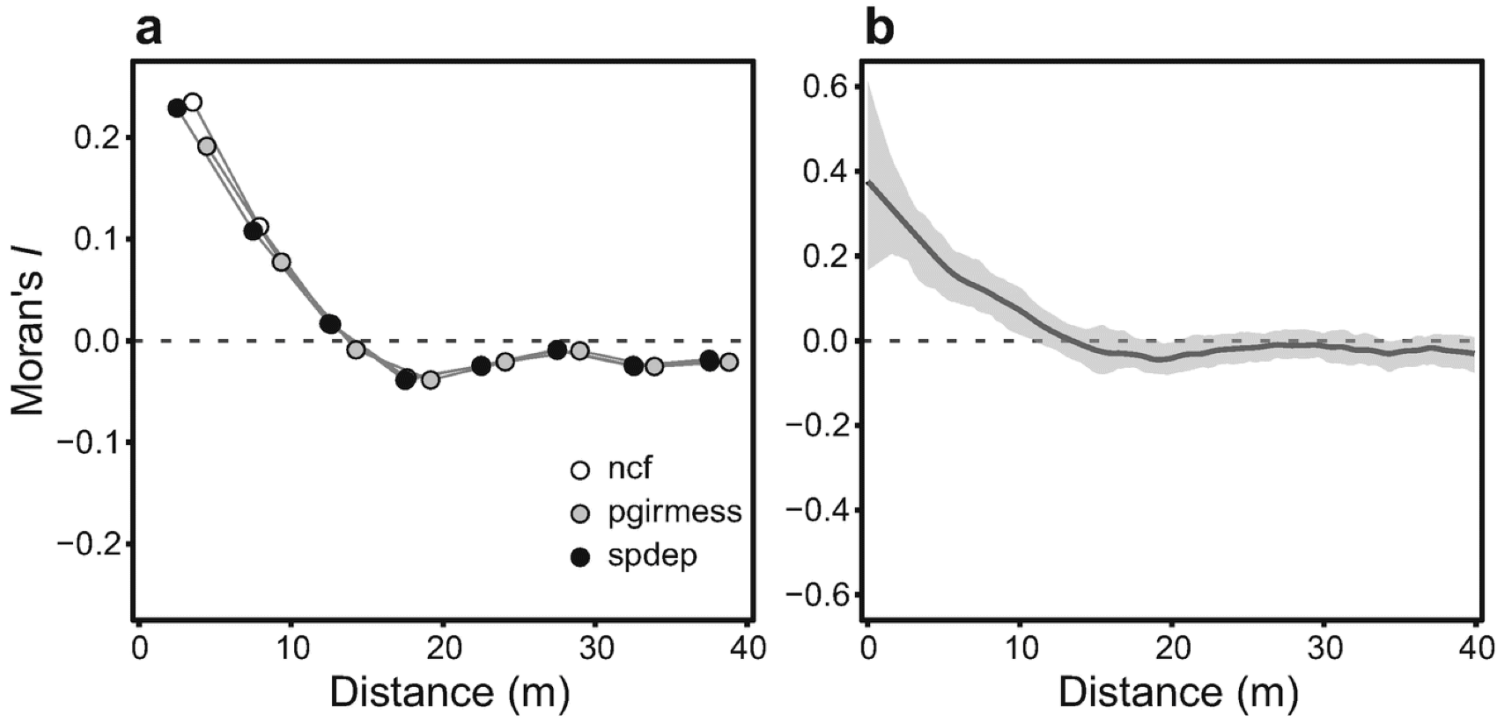
- We can quantify Moran's I for different neighborhoods.
- Define distance range, d, for 'ring' style neighborhoods.
  - Reminiscent of the pair correlation function
- Plot of I vs. distance is a **correlogram**.

Correlograms

We can calculate I for different distances, or distance classes, between points by using different weight matrices.

If we make a plot $I(d)$ vs d we get a correlogram.

- How do you interpret these correlograms?



Quantifying Autocorrelation: Semivariance and variograms

Semivariance and variograms

The semivariance equation looks complicated:

$$\gamma(d) = \frac{1}{2n(d)} \sum_{i=1}^{n(\mathrm{d})} [z(x_i + d) - z(x_i)]^2$$

- $z(x_i)$ and $z(x_i + d)$: values of a pair of points separated by distance d
- $n(d)$ is the number of points separated by distance d

Stochastic Processes

Stochastic processes and inference

- What is a stochastic processes?
- What is a realization of a stochastic process?
- How does it relate to model thinking and the dual-model paradigm?

“Inference requires many realizations…”

- Oliver, M.A., and Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. CATENA 113, 56–69.

Semivariance and Variograms

The semivariance equation:

$$\gamma(d) = \frac{1}{2n(d)} \sum_{i=1}^{n(h)} [z(x_i + d) - z(x_i)]^2$$

is hard to decipher, but we can break it down into sensible components, starting with the function $z(\,)$.

In probability theory terms, $z(x_i)$ is a *realization* of a *random process*, Z, that occurs somewhere in space. $z(x_i)$ is an observation at location i.

Deconstructing Semivariance: Point Pairs

The component:

$$\sum_{i=1}^{n(h)} [z(x_i + d) - z(x_i)]^2$$

symbolically represents pairs of points. It's analogous to this component of Moran's I:

$$\frac{\sum_i \sum_j (x_i - \bar{x})\left(x_j - \bar{x}\right)}{N_{pairs}}$$

Semivariance and variograms

The semivariance equation looks complicated:

$$\gamma(d) = \frac{1}{2n(d)} \sum_{i=1}^{n(\mathrm{d})} [z(x_i + d) - z(x_i)]^2$$

Specific formula details aren't super important for us to remember, but we should recognize the similarity to a variance formula.

- Semivariance: sum of differences in values of points separated by a constant distance.
- Note that in practice we use a distance class, or distance range.

Semivariance Intuition

Recall that semivariance is a measure of variability at a fixed distance, or distance class/bin.

With positive autocorrelation/dependence, we expect:

- Autocorrelation (as measured by I) to be high at short distances.
- Semivariance to be low at short distances.

How do we expect a plot of semivariance to behave under different scenarios?

Variograms

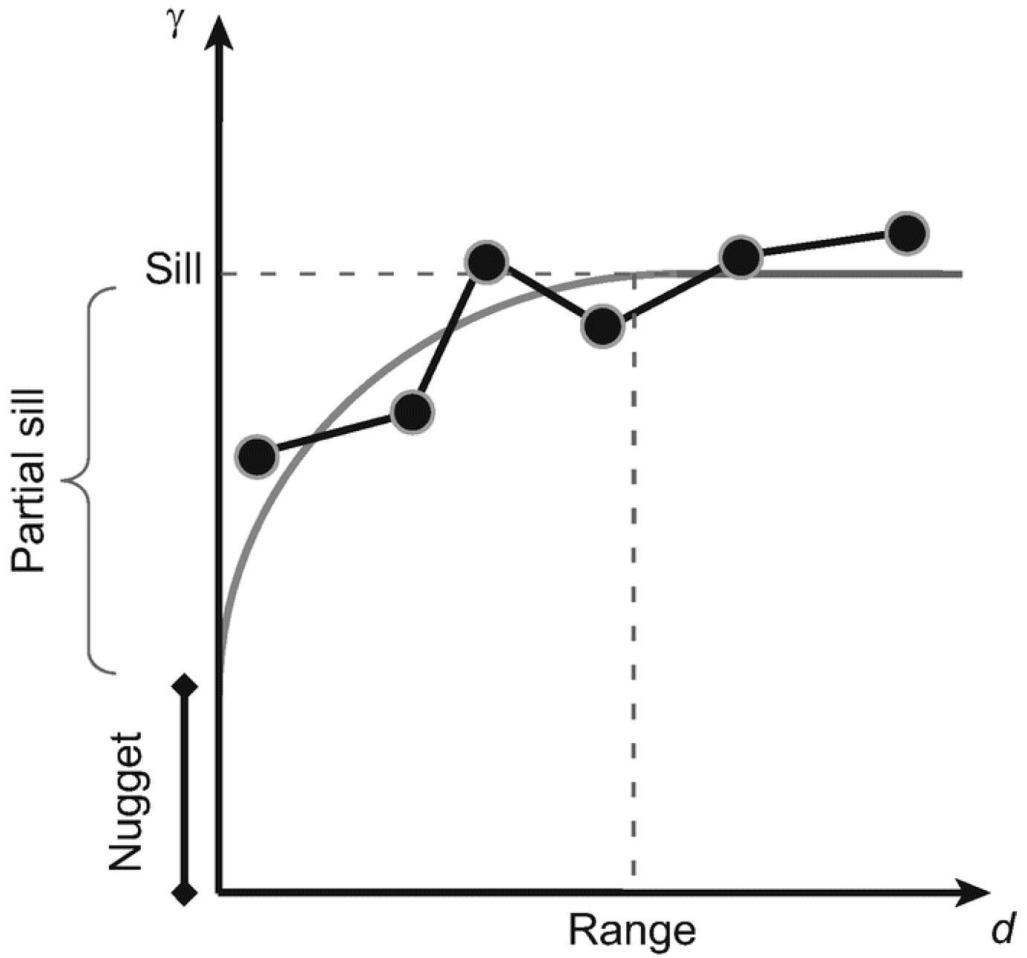
Variograms are plots of γ versus distance.

What could we learn from a variogram?

Variograms are associated with a fun set of terms.

- sill
- nugget
- range

- F+F figure 5.2



Variograms

Nugget: amount of variation at short distances.

- Amount of variation with local influence
- Background noise, measurement error, unmeasured variables

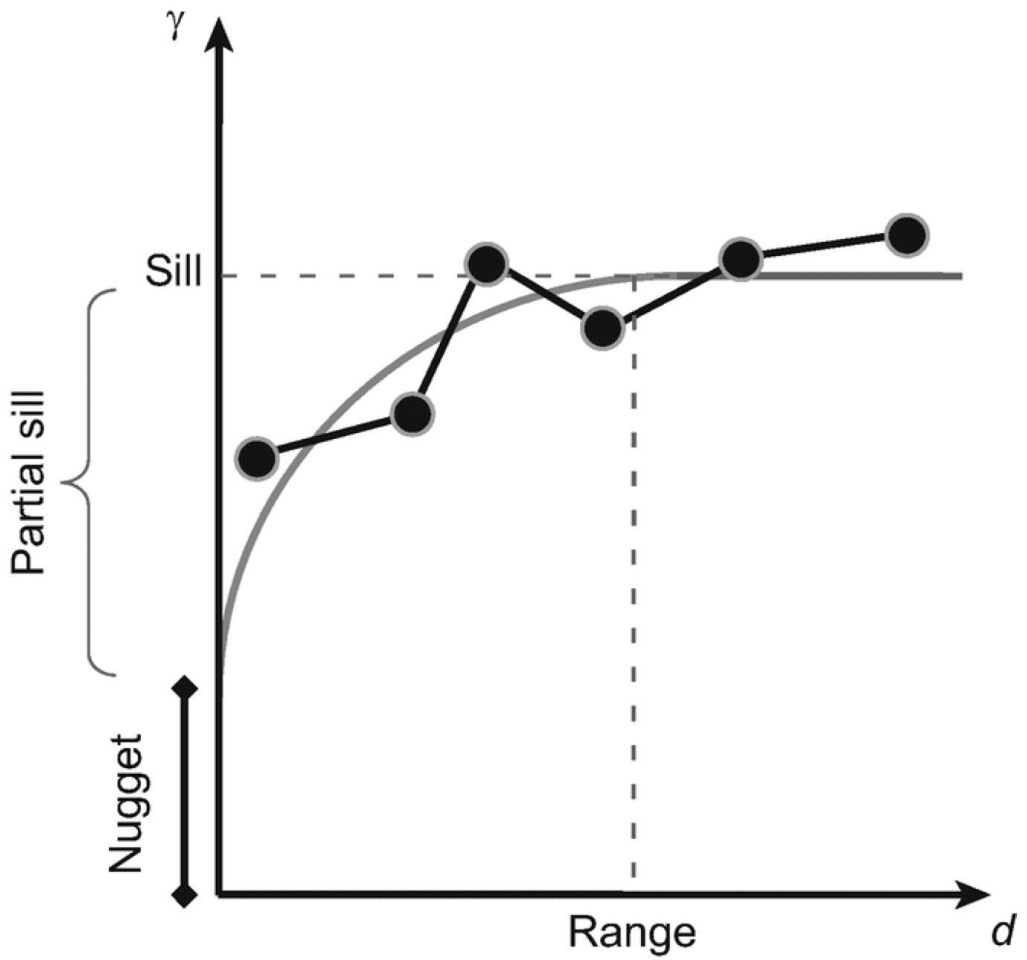
Range: distance above which there is no spatial dependence

- Points separated by this distance are no longer autocorelated

Sill: Amount of variation at long distances.

- Amount of variation without local influence.

- F+F figure 5.2



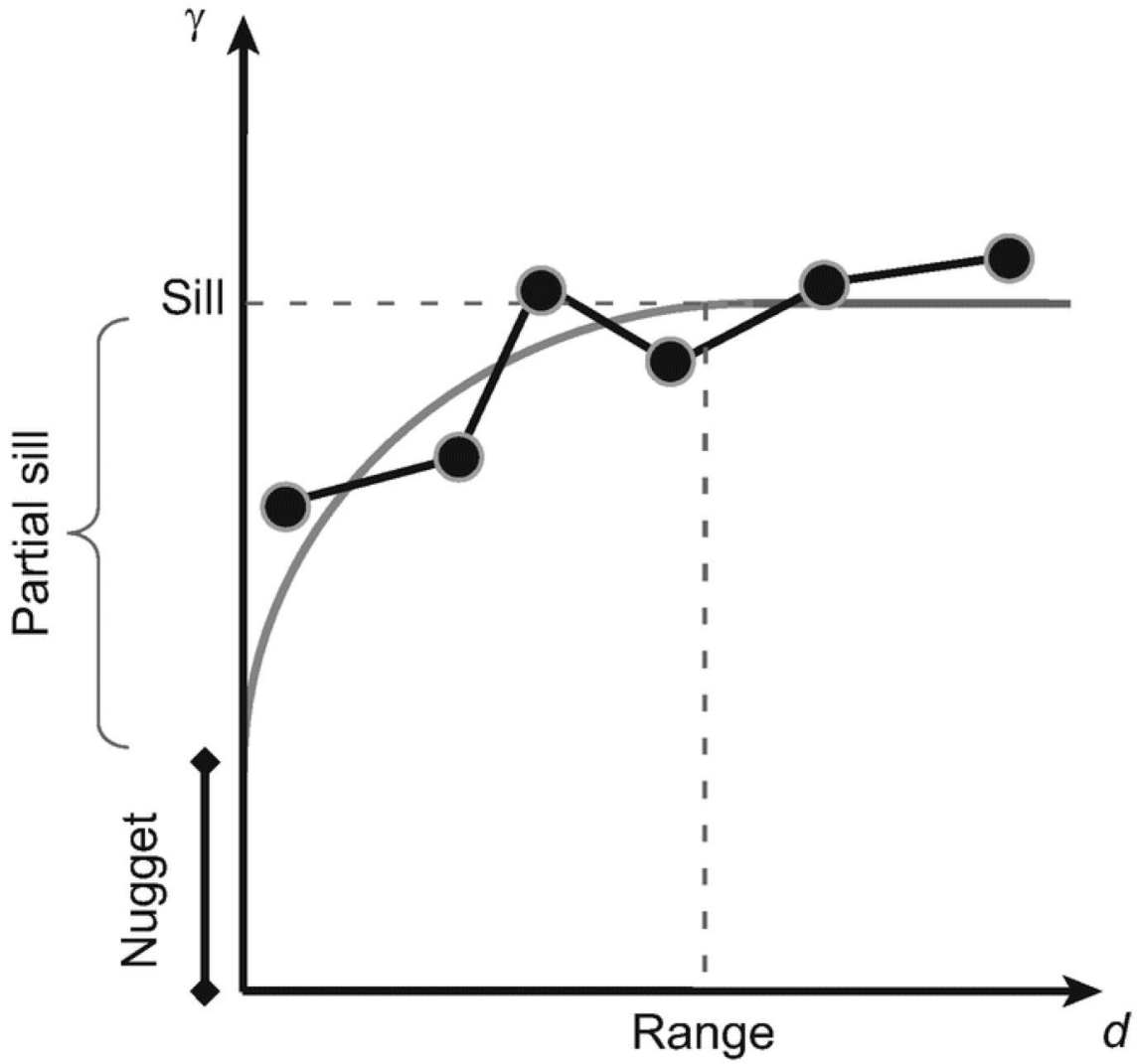
Spatial Statistics and Geostatistics

- Moran's I: spatial form of Pearson's correlation
- Semivariance (γ): spatial form of variance
- Interpolation methods:
  - nearest-neighbor interpolation
  - inverse distance weighted interpolation
  - **Kriging**
- A correlogram is a plot of $I(d)$.
- A variogram is a plot of $\gamma(d)$

Variogram Components: The Nugget

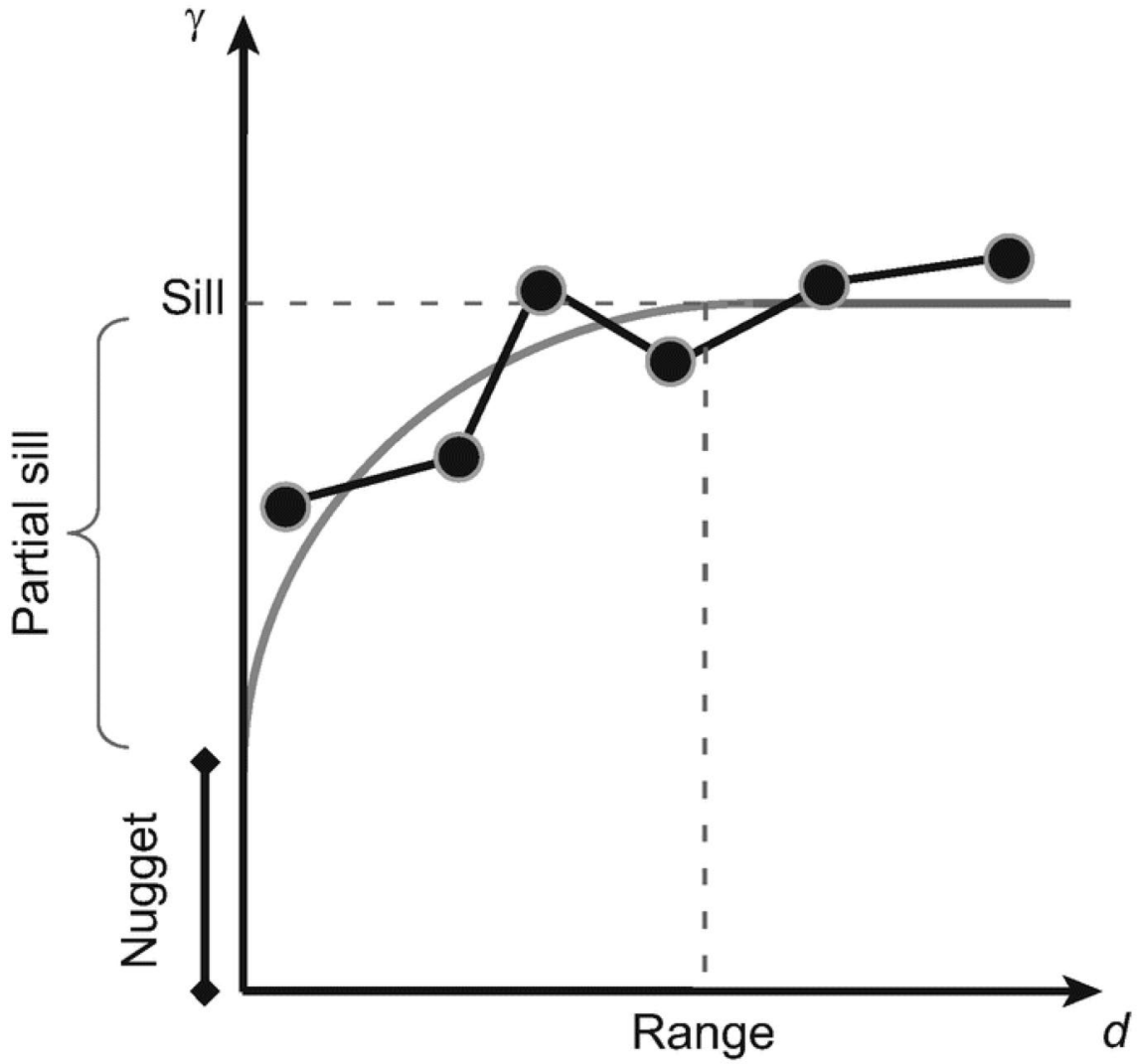
How much variability do we expect at very nearby points?

Can a nugget ever be zero?



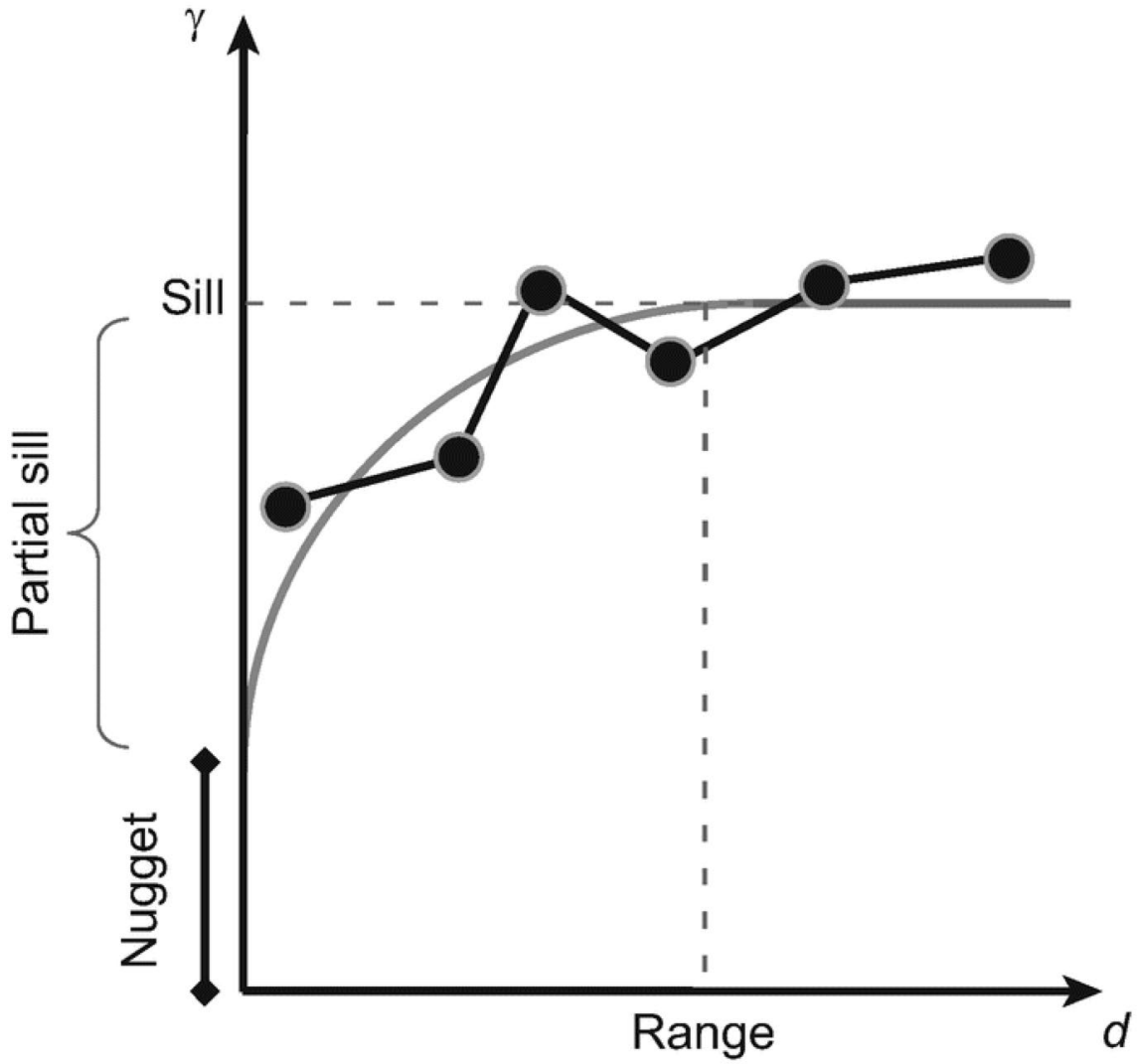
Variogram Components: The Range

What happens as we increase the distance between pairs of points?



Variogram Components: The Sill

What happens between pairs of points separated by large distances?



Anatomy of a Variogram

The variogram components help us answer different questions:

- nugget: How much variability is not explained by spatial proximity?
- range: How far do points have to be separated for spatial dependence to break down?
- sill: What is the variability among points that are distant enough to be spatially independent?

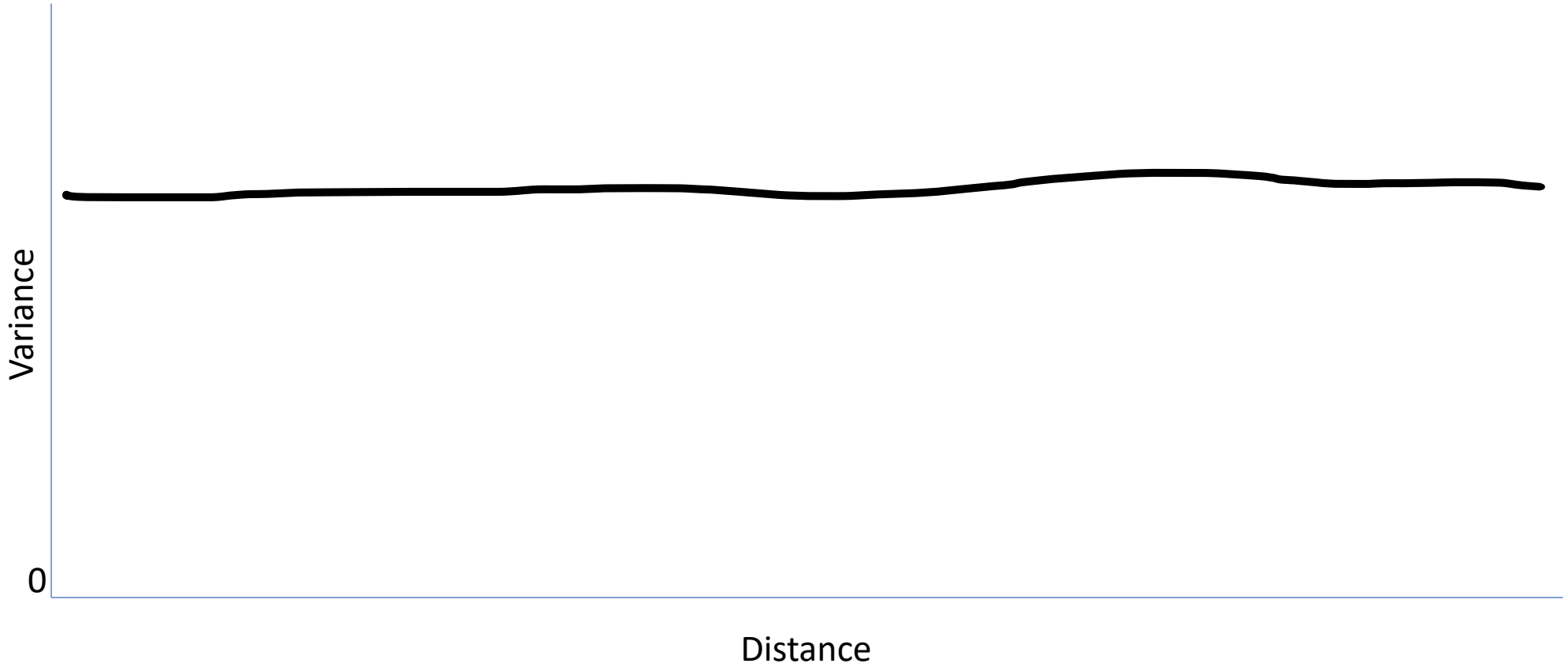
Sketching Variograms and Correlograms

Let's draw some variograms and correlograms:

- We have observed a value of Z at one location.
  - Scenario 1: Knowing the outcome of a stochastic process, z_i, tells us nothing about any other realizations (nearby or far)
  - Scenario 2: Nearby points are similar, separated points are different.
  - Scenario 3: No spatial dependence.
  - Scenario 4: Nearby points are identical, far points are not correlated

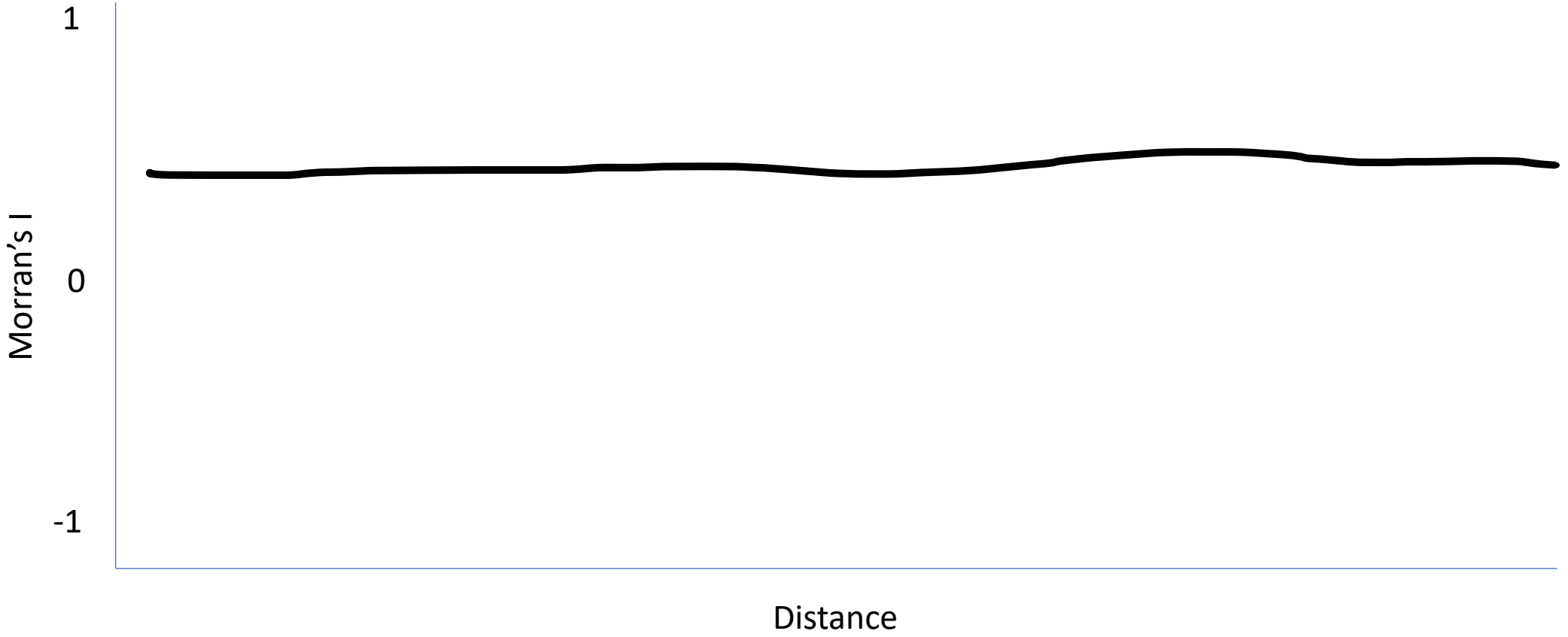
Variogram 1

- Scenario 1: Knowing the outcome of a stochastic process, z_i, tells us nothing about any other realizations (nearby or far)



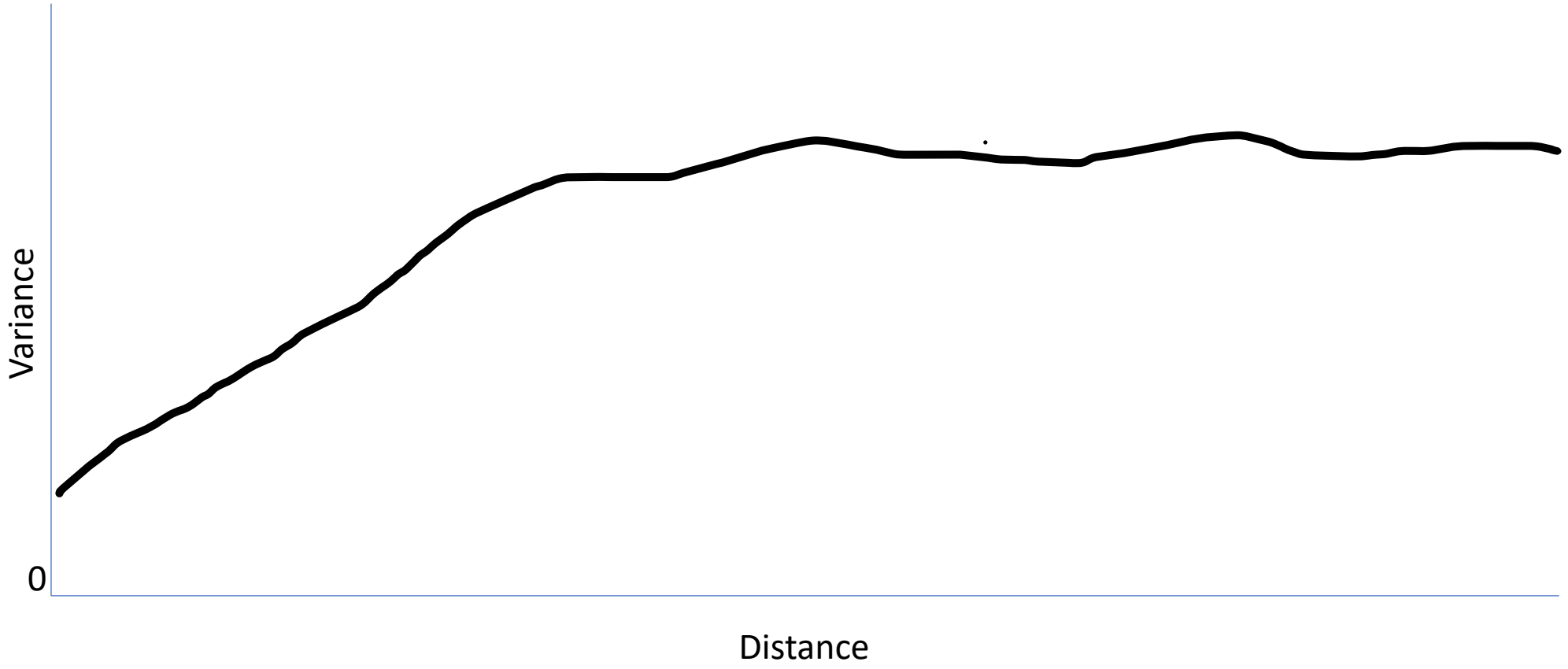
Correlogram 1

- Scenario 1: Knowing the outcome of a stochastic process, z_i, tells us nothing about any other realizations (nearby or far)



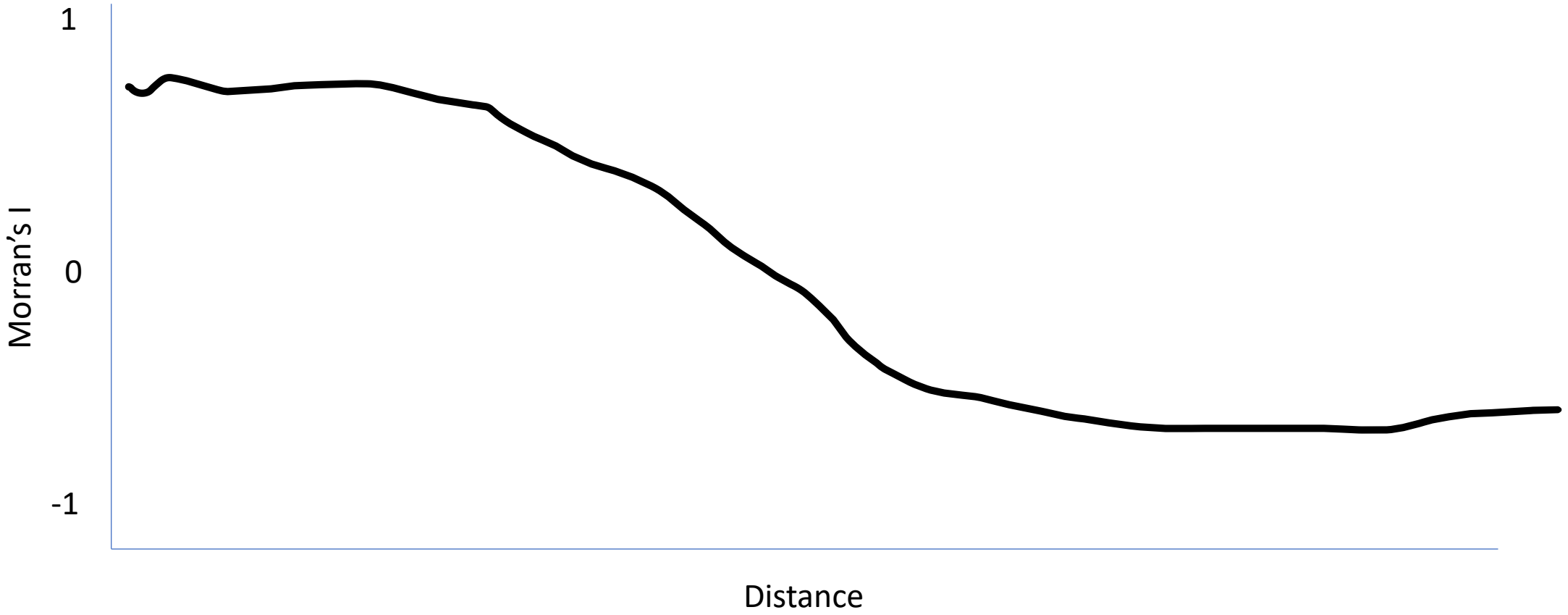
Variograms 2

- Scenario 2: Nearby points are similar, separated points are different.



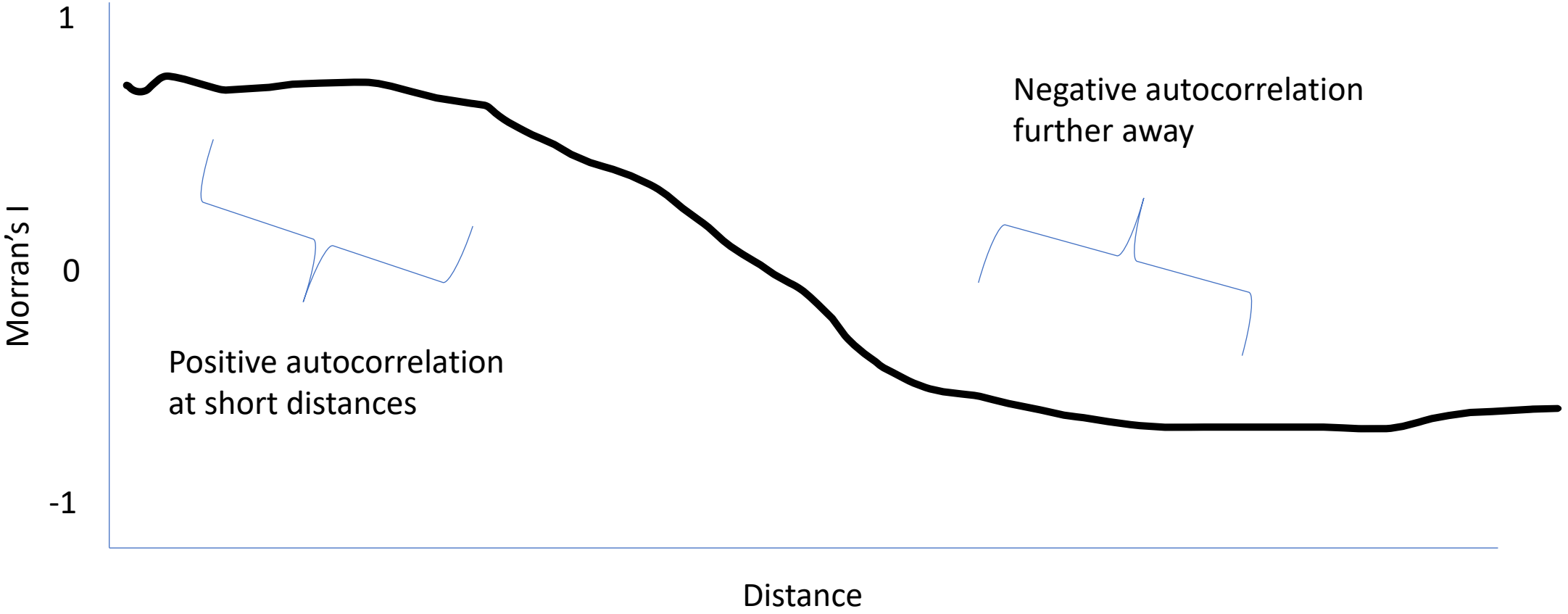
Correlogram 1

- Scenario 2: Nearby points are similar, separated points are different.



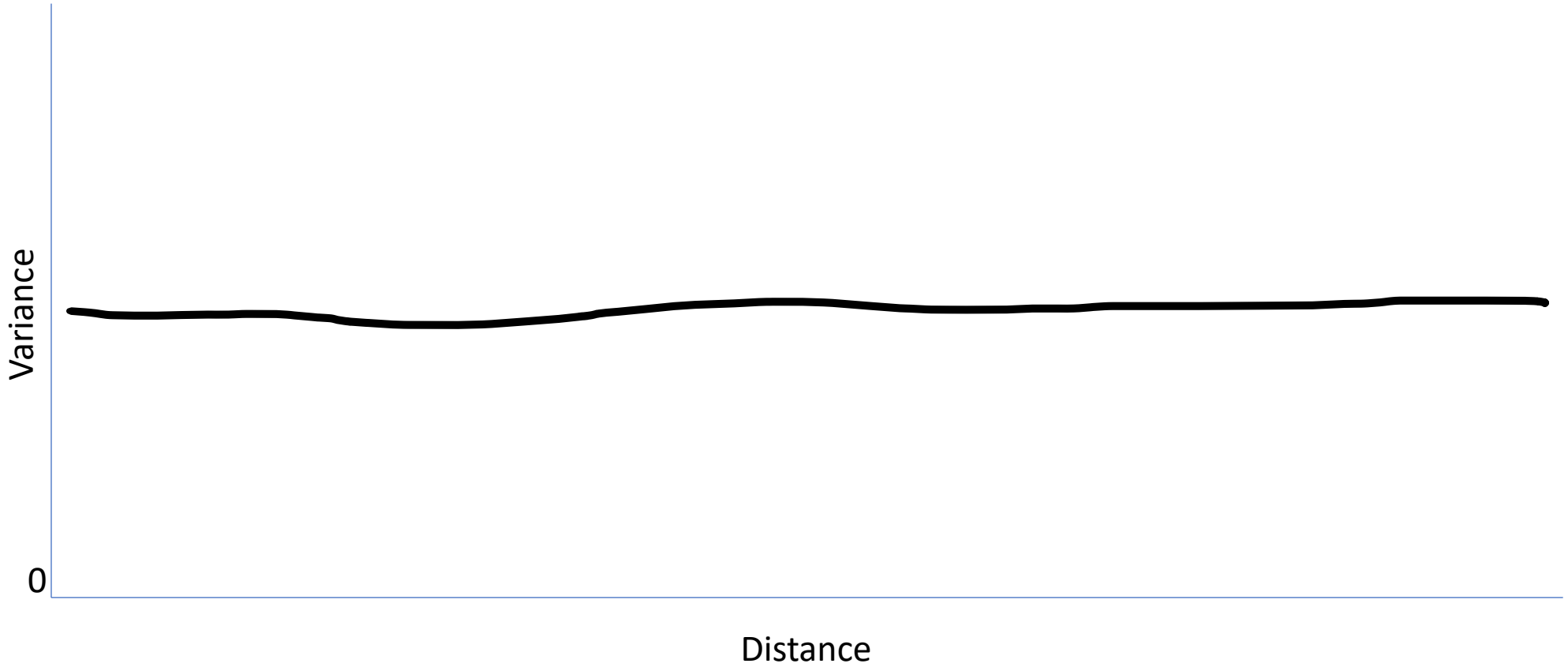
Correlogram 1

- Scenario 2: Nearby points are similar, separated points are different: kind of like overdispersion



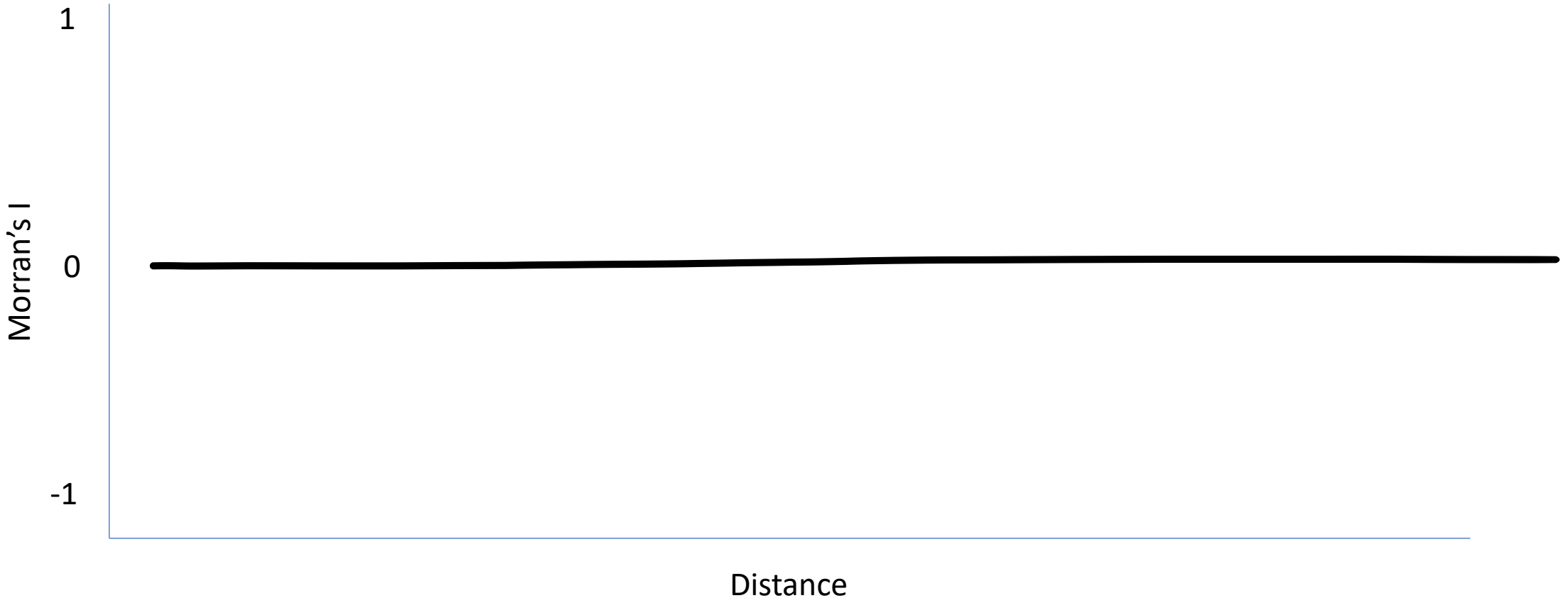
Variogram 3

- Scenario 3: No spatial dependence.



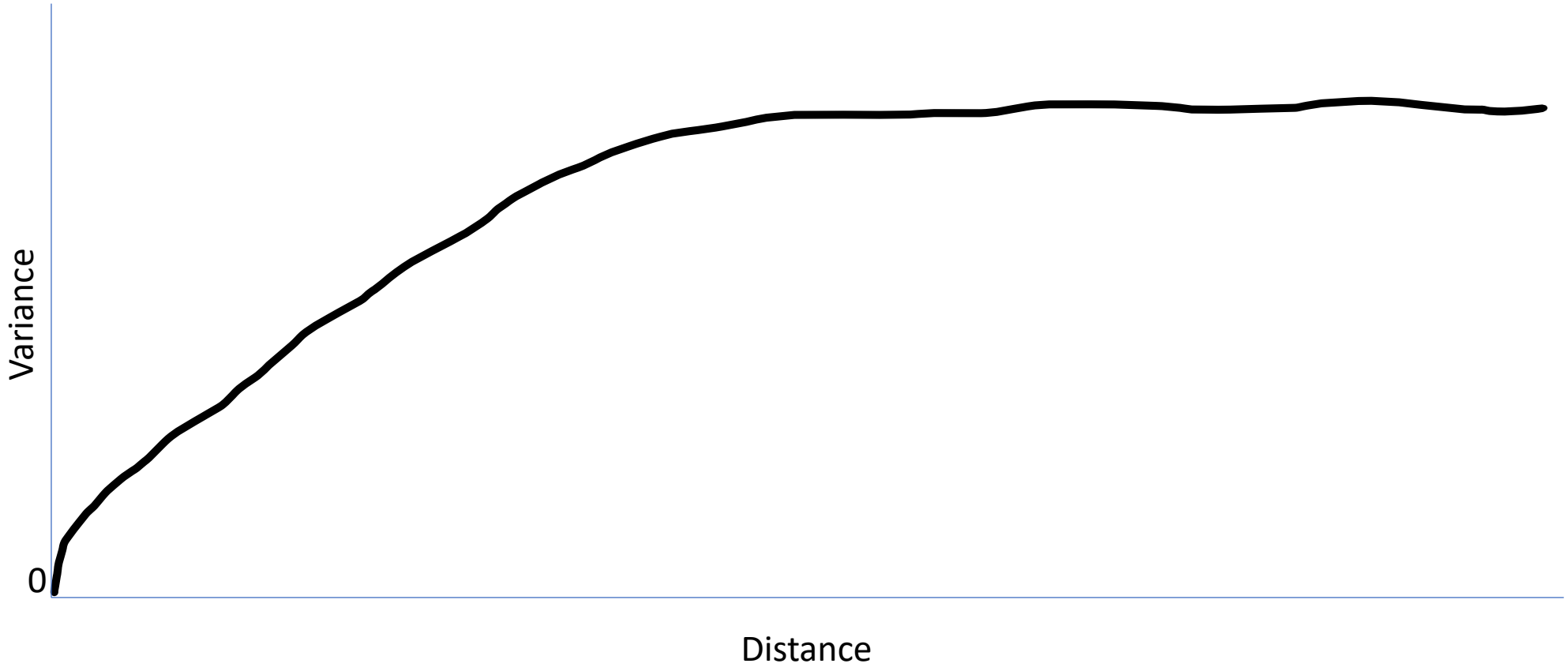
Correlogram 3

- Scenario 3: No spatial dependence.



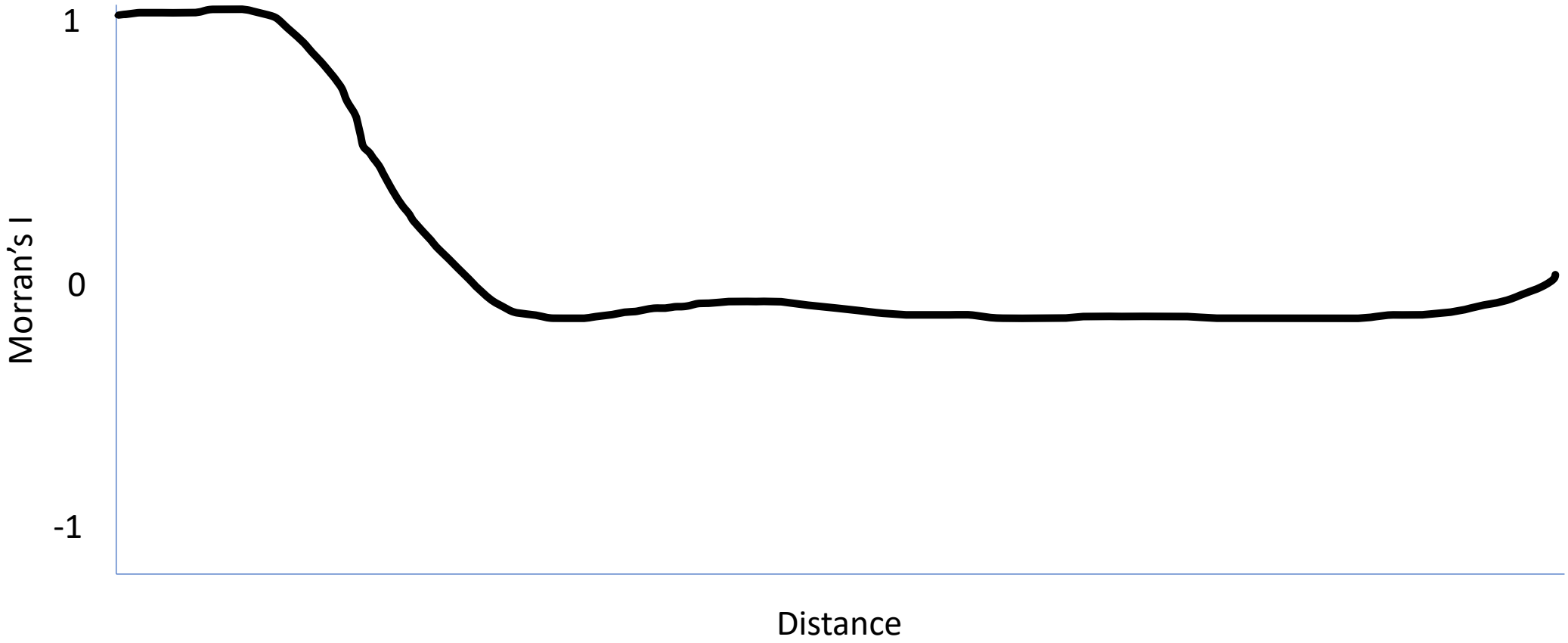
Sketching Variograms

- Scenario 4: Nearby points are identical, far points are not correlated



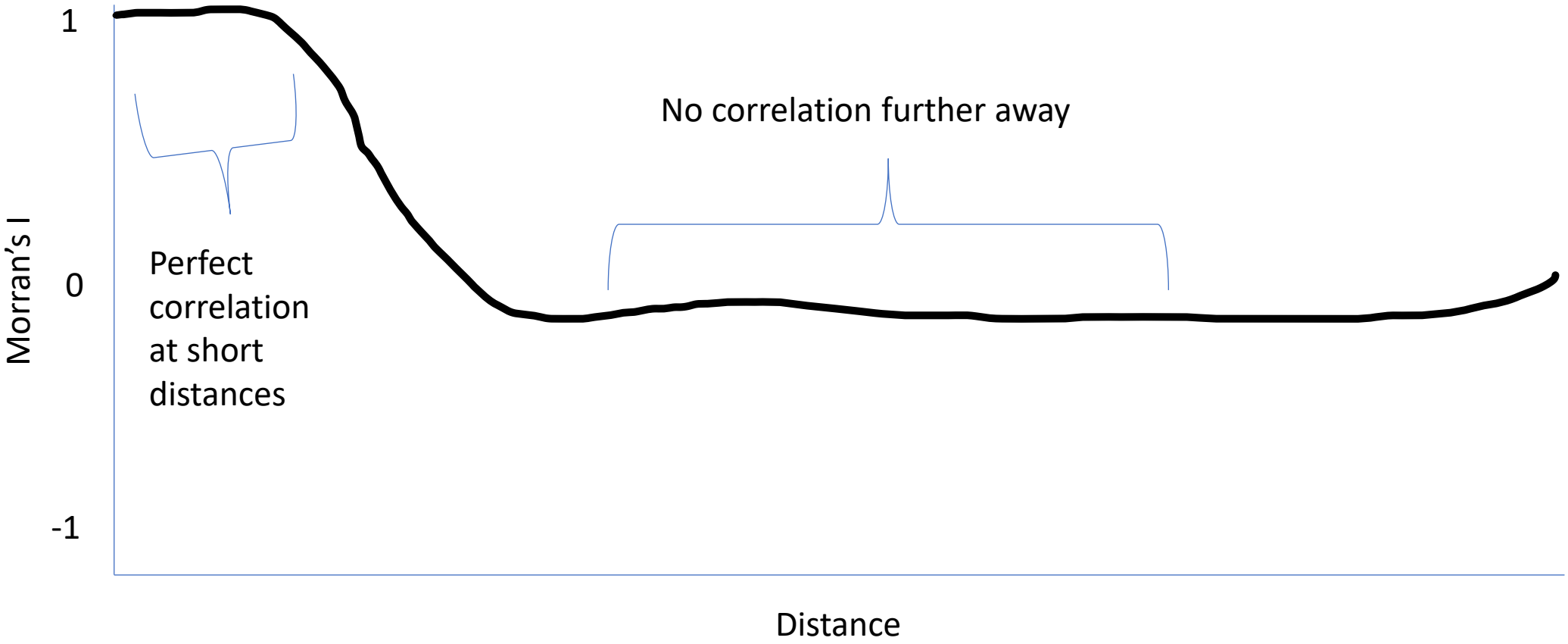
Correlogram 4

- Scenario 4: Nearby points are identical, far points are not correlated



Correlogram 4

- Scenario 4: Nearby points are identical, far points are not correlated



Moran's I and Semivariance

We can think of the differences with the help of the following informal definitions:

$I(d)$: "correlation as a function of distance"

$\gamma(d)$: "variance as a function of distance"