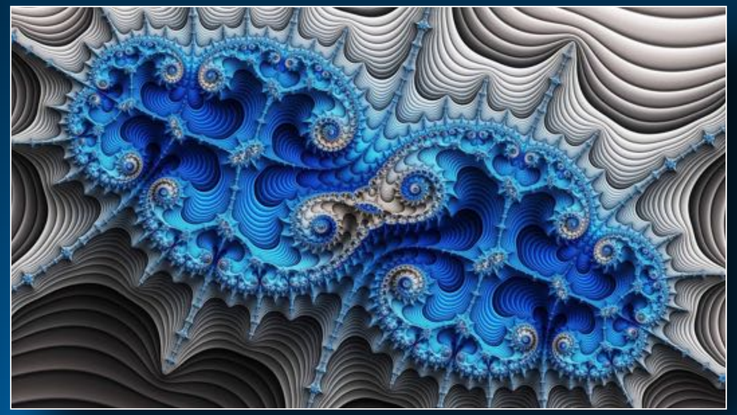
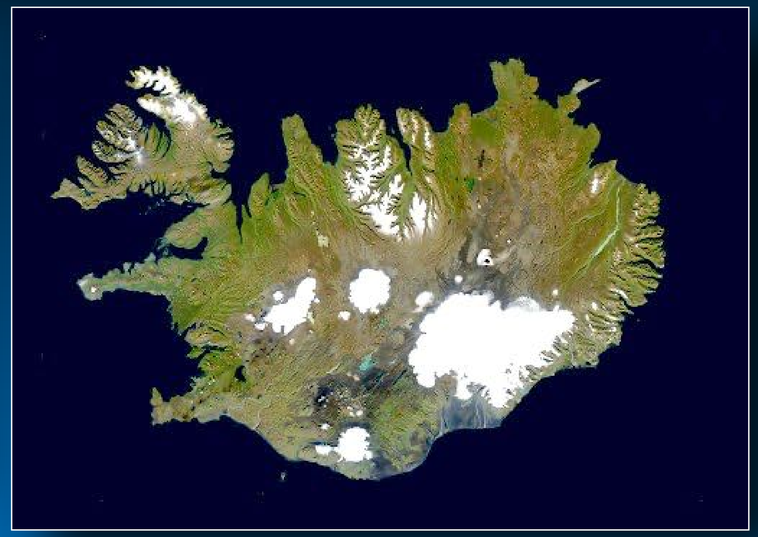
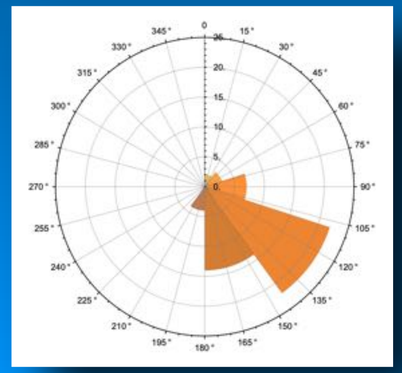
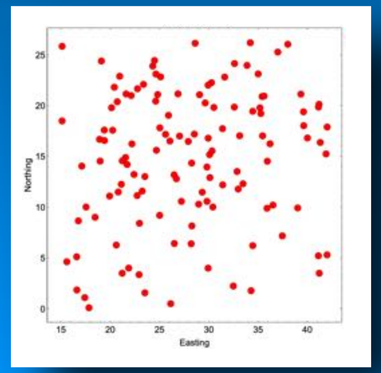
Spatial & Directional Data Analysis

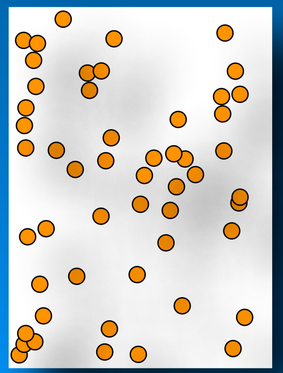
Prof. Norman MacLeod
School of Earth Sciences & Engineering, Nanjing University





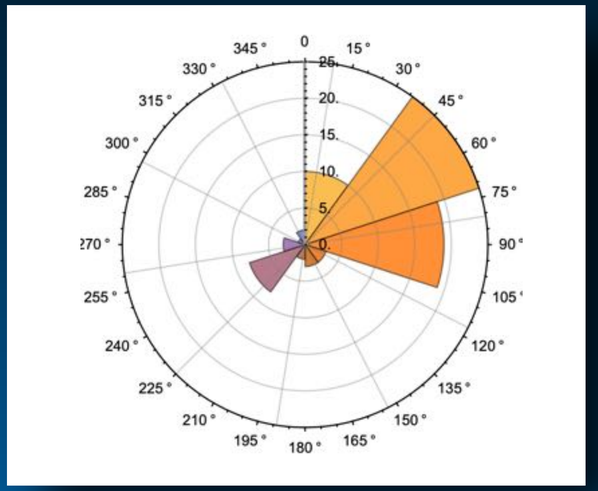


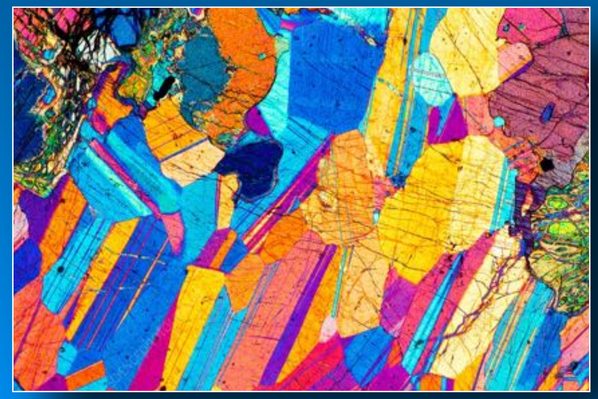




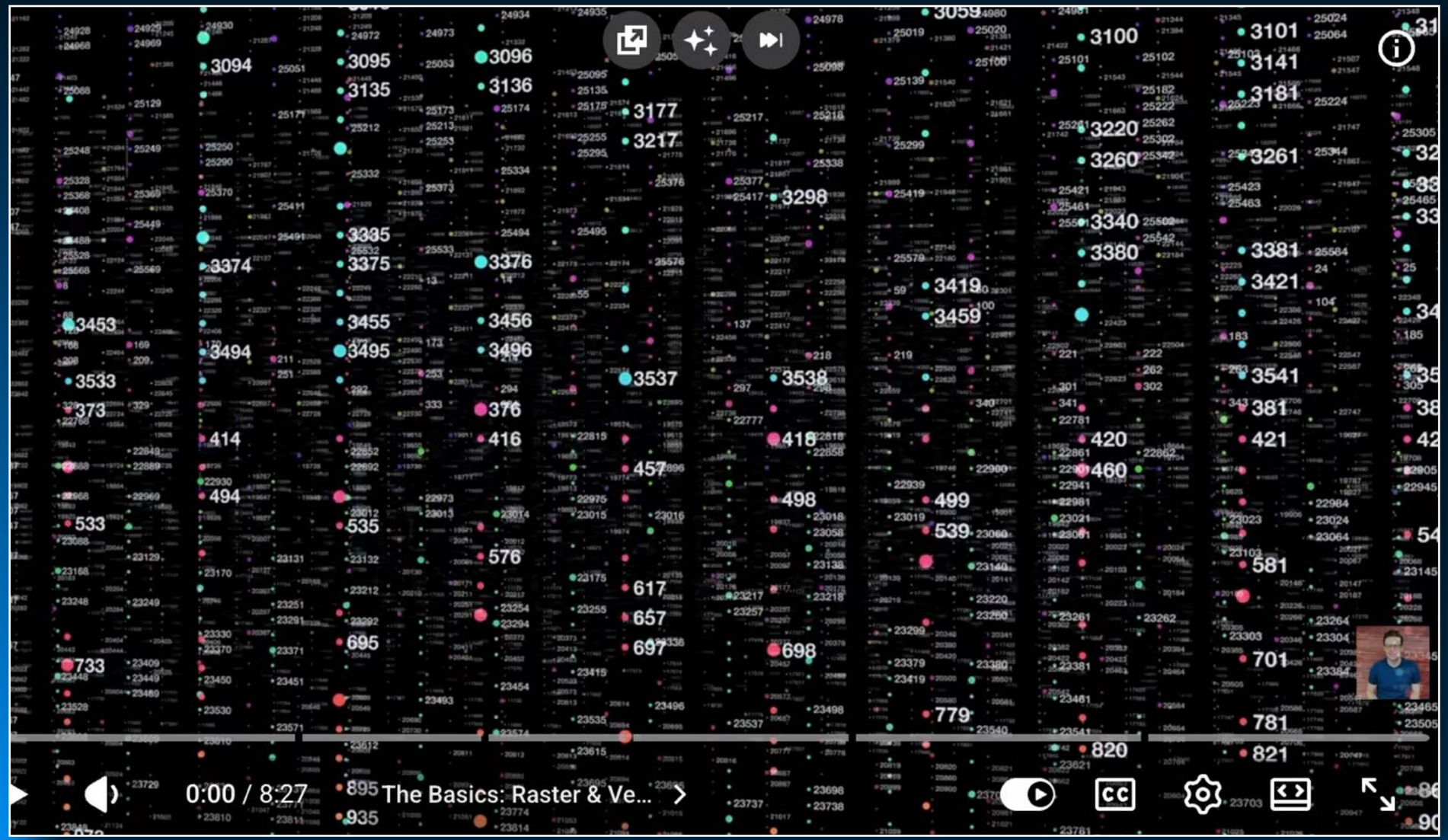
Spatial Data

Examples of Spatial Data in Earth Science





Spatial Data



Spatial Data

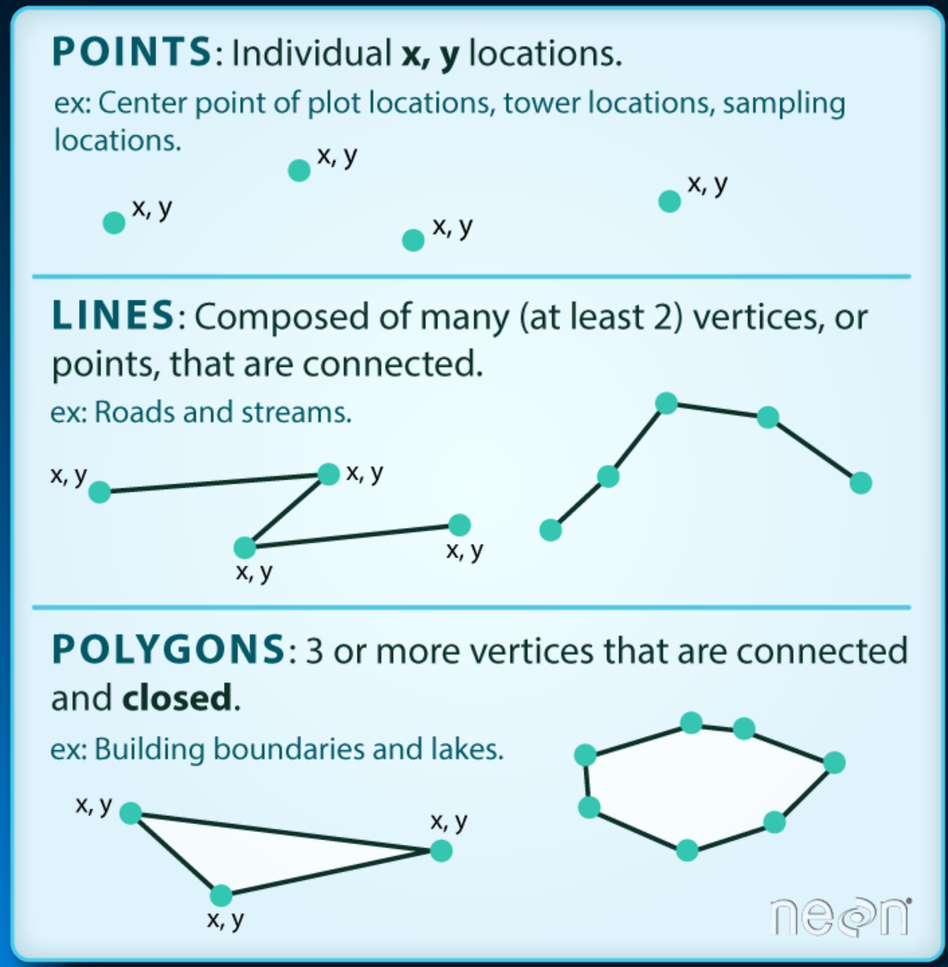
- Data that specify the locations of objects and/or features in a two or three-dimensional geometric space (2D or 3D respectively).
- Geospatial data are a particularly diverse sub-set of spatial data whose scope encompasses locations and features present on the Earth and other planets. But spatial data can encompass other aspects of geometric analysis that are restricted to more localized domains (e.g., patterns exhibited by thin sections or electron micrographs, distributions of points in ordination spaces).

- Some make a distinction between spatial and attribute data, but in reality these are simply different types of spatial data. All can be analyzed quantitatively and there are many benefits that proceed from adopting this approach.

Types of Spatial Data

Vector Data
Locations of discrete objects and well-defined locations using mathematical points, lines and polygons.

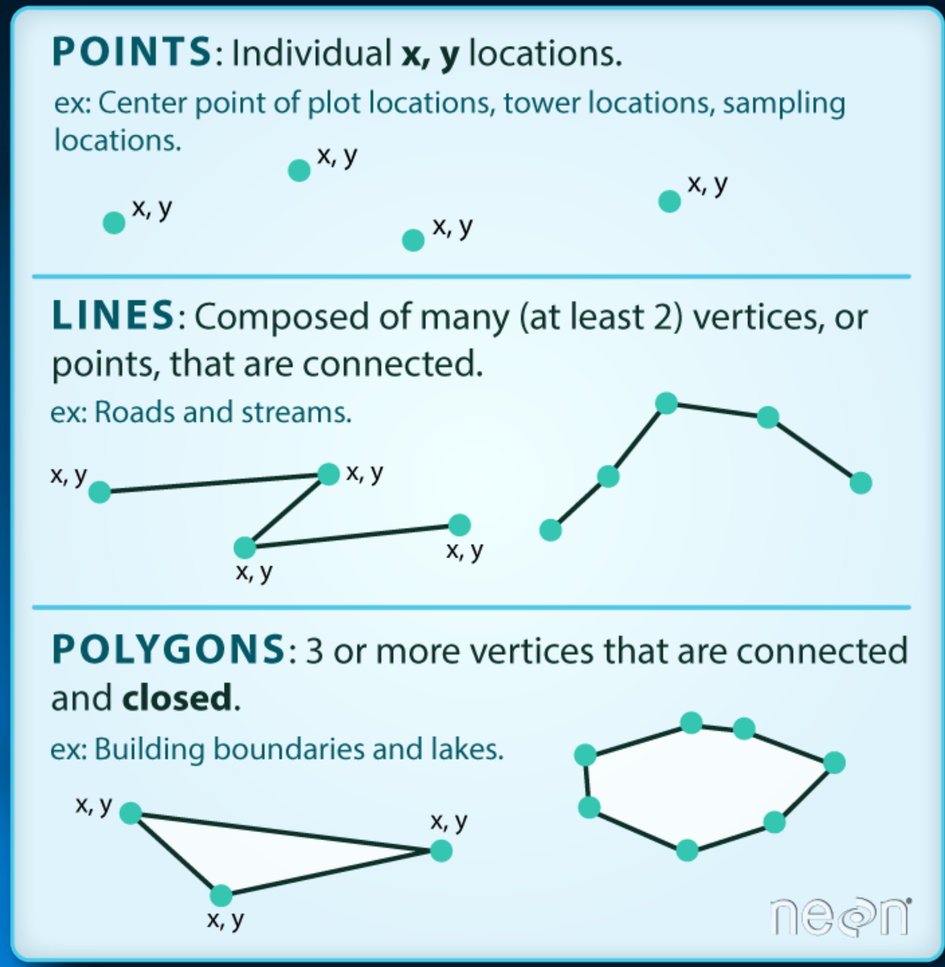
- Points - specific locations without size or shape.
- Lines/Polylines - linear features connecting multiple points
- Polygons - enclosed areas with well-defined boundaries.



Types of Spatial Data

Advantages of Vector Data

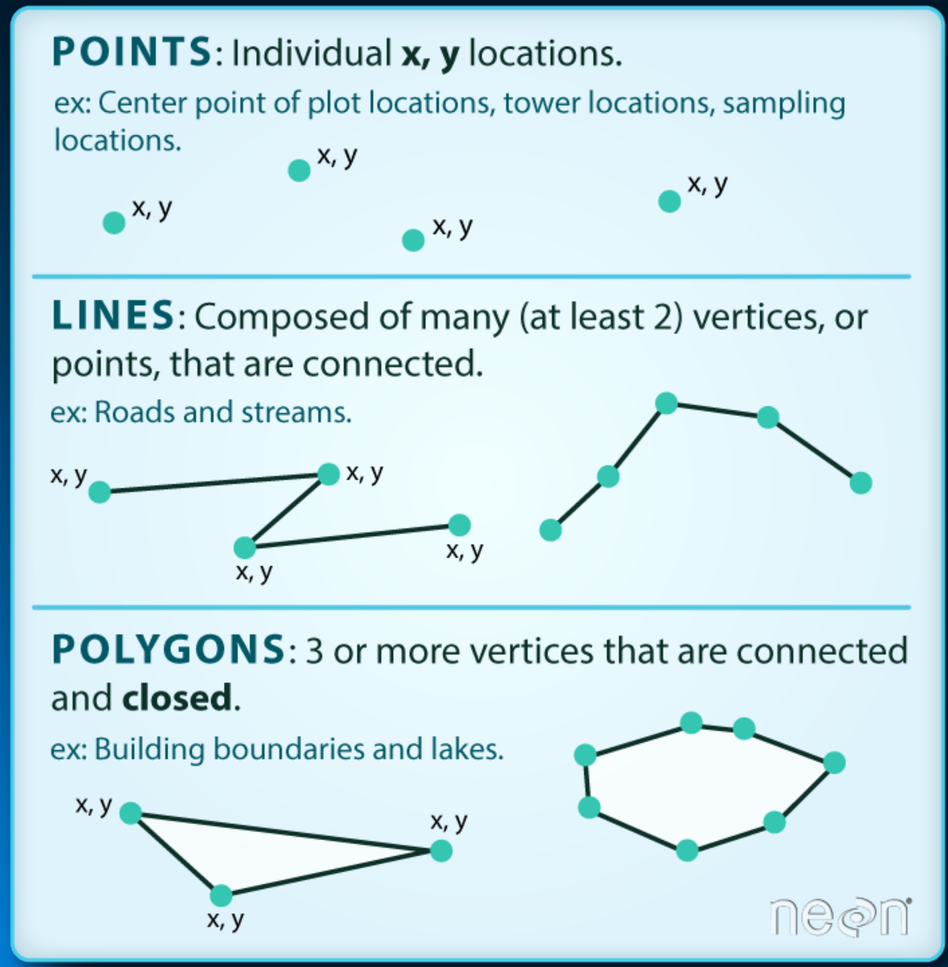
- The geometry itself contains information about what the dataset creator thought was important.
- The geometric structures hold information in themselves - why choose point over polygon, for instance?
- Each geometric feature can carry multiple information attributes (e.g., a database of cities can have attributes for name, country, population).
- Data storage can be very efficient compared to raster data.



Types of Spatial Data

Disadvantages of Vector Data

- Potential loss of detail compared to raster data.
- Potential bias in datasets - what didn't get recorded?
- Calculations involving multiple vector layers need to analyze patterns in the geometry of the data as well that of the vector attributes. As a result data analysis can be slow compared to the analysis of raster data.



Types of Spatial Data

Raster Data
A grid of pixels (cells), where each pixel stores a value representing a specific geographic attribute. This format is ideal for representing continuous surfaces, environmental data, and imagery.

- Satellite Imagery - provides large-scale views of Earth's surface, making it essential for environmental monitoring, land-use analysis, and disaster response.
- Aerial Photography - captures high-resolution images from aircraft or drones, offering detailed visual data for applications such as urban mapping, disaster assessment, and forestry analysis.
- Digital Elevation Models (DEM) - represent terrain elevations and landforms, commonly used in 3D terrain modeling, flood-risk analysis, and mountain topography studies.

Types of Spatial Data

Advantages of Raster Data

- Can contain multiple levels (bands) of information.
- Facilitates representation as continuous surfaces.
- Potentially very high levels of detail.
- Data are ‘unweighted’ across their extent - the geometry doesn’t highlight features implicitly.
- Cell-by-cell calculations can be very fast and efficient.

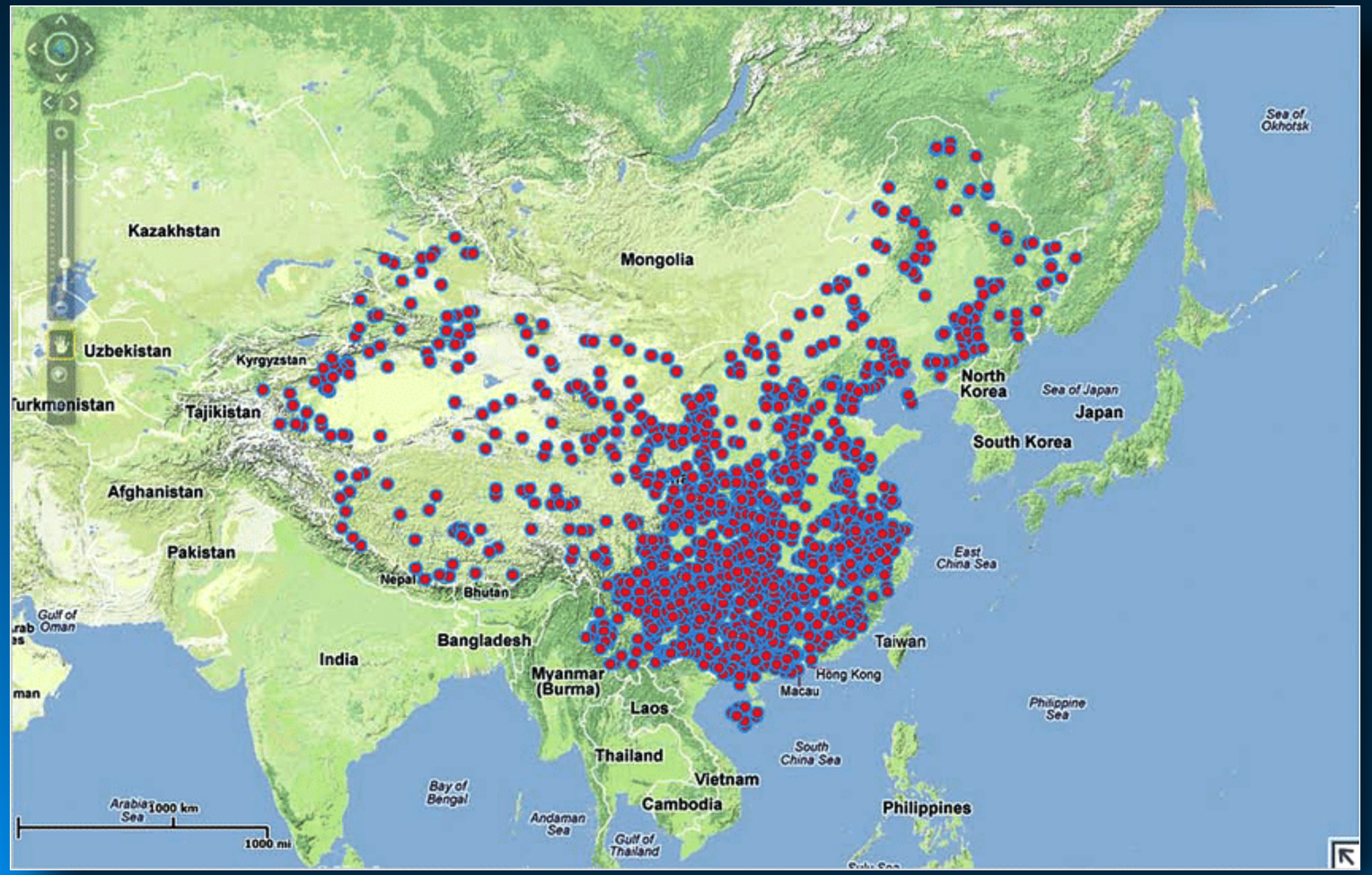
Types of Spatial Data

Disadvantages of Raster Data

- Very large file sizes are produced as the raster-cell size gets smaller.
- Popular formats currently don’t embed meta-data well.
- Can be difficult to represent complex information.

Vector Data

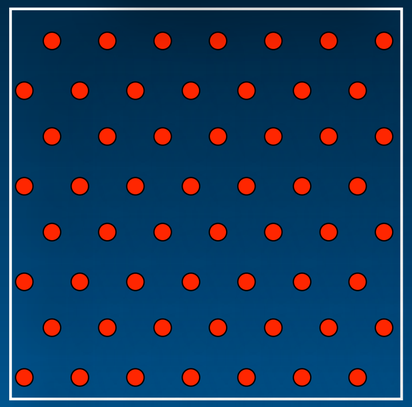
Example Analyses: Distribution of Points



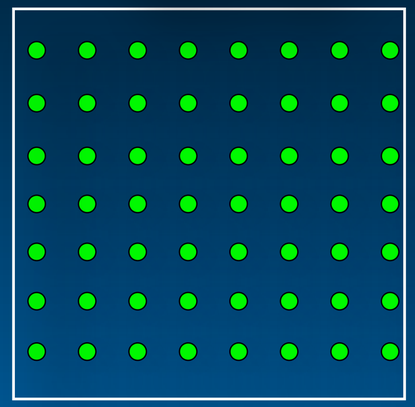
Distribution of Points

The concept of the point is fundamental to spatial data.

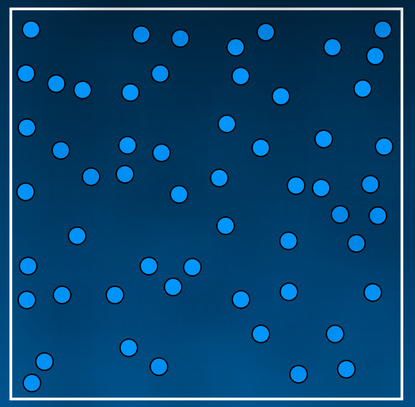
Regular Spacing,
Hexagonal Grid



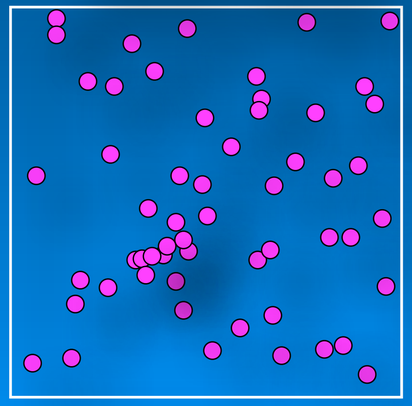
Regular Spacing,
Square Grid



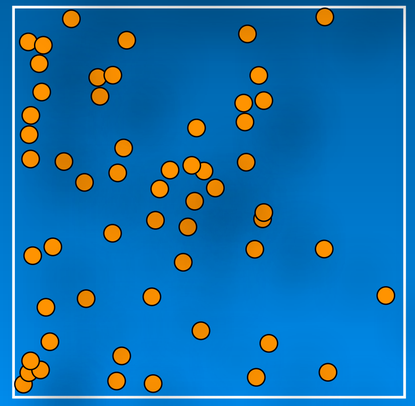
Random Spacing
within a Square Grid



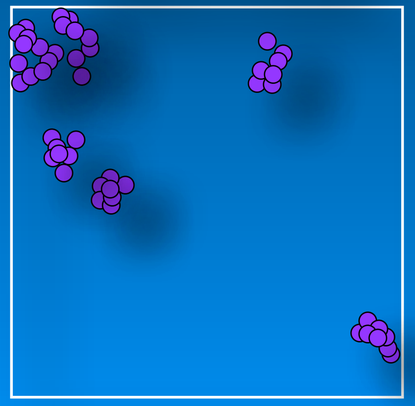
Bivariate, Uniform
Random Spacing



Non-Uniform Spacing.
Log Scaling on x-Axis



Random Distribution
About 8 Centers



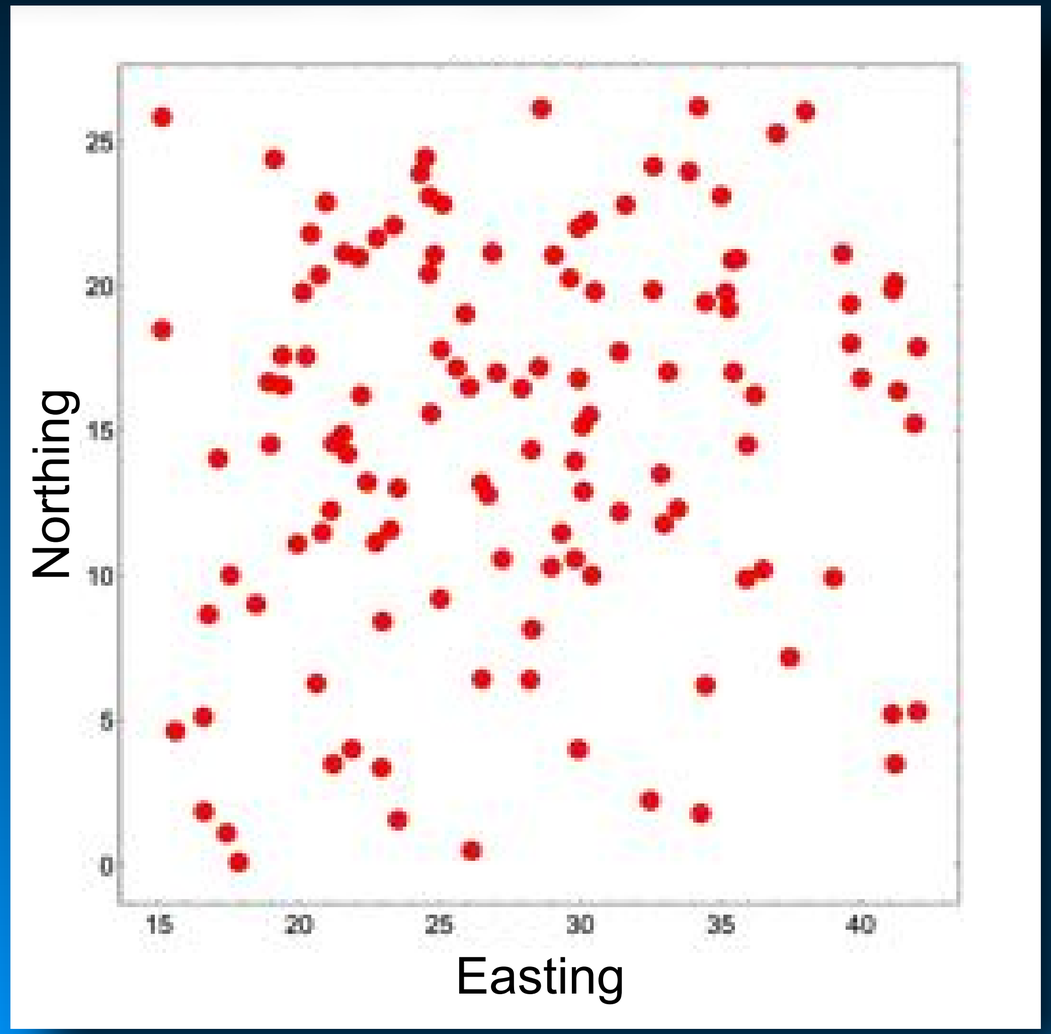
Distribution of Points

Test for Uniform Distribution of Points

Often one of the first questions to be confronted in an earth science context is what sort of spatial distribution the sampled localities in a dataset exhibit.

This is a plot of the locations of 123 wells drilled in the Arbuckle Group in central Kansas.

Do these data exhibit uniform density?



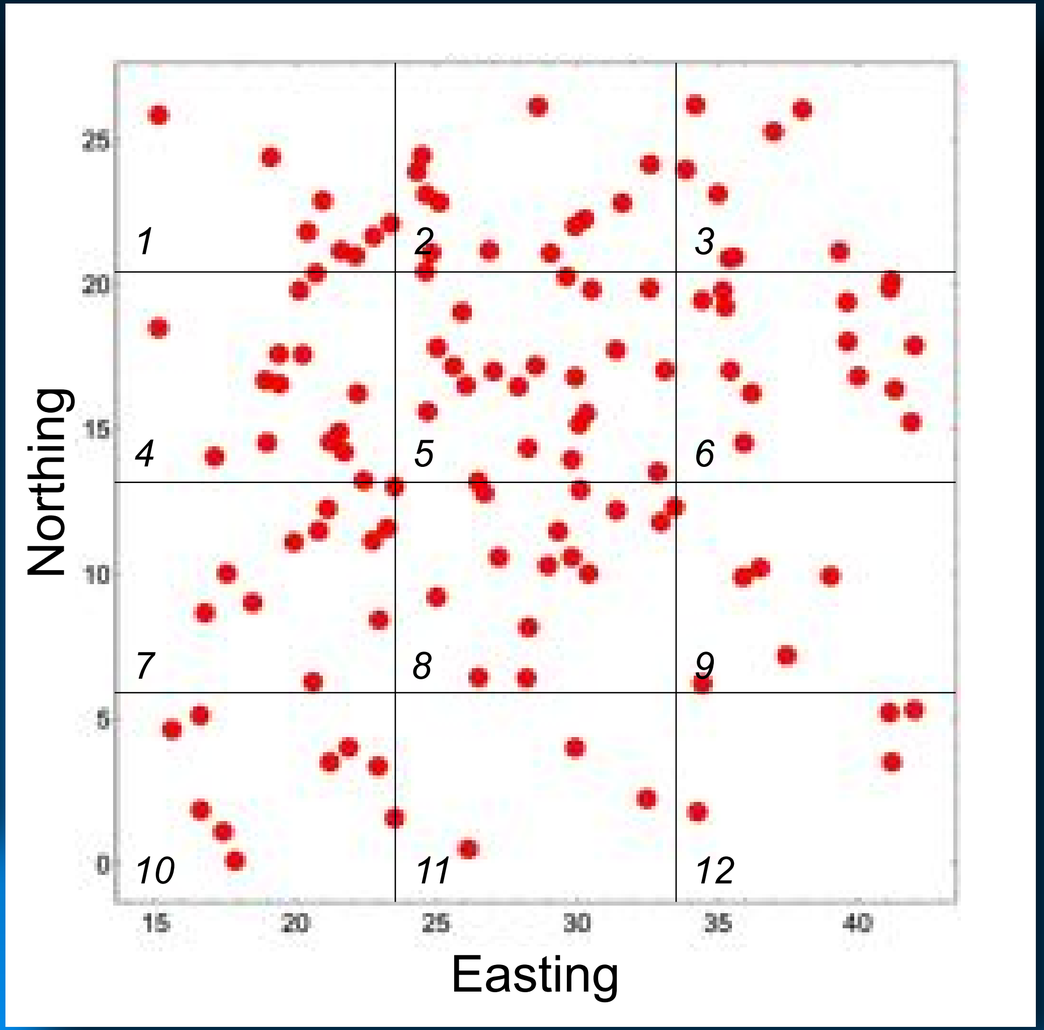
Distribution of Points

Test for Uniform Distribution of Points: Quadrat Analysis

One approach would be to partition the area into a series of equal-size quadrats and count the frequency of wells that fall into each partition.

If the wells are spaced uniformly there should be no statistically significant difference between the quadrat frequency counts.

Cell	Obs.
1	8
2	13
3	8
4	14
5	20
6	14
7	11
8	14
9	5
10	9
11	3
12	4
Σ	123

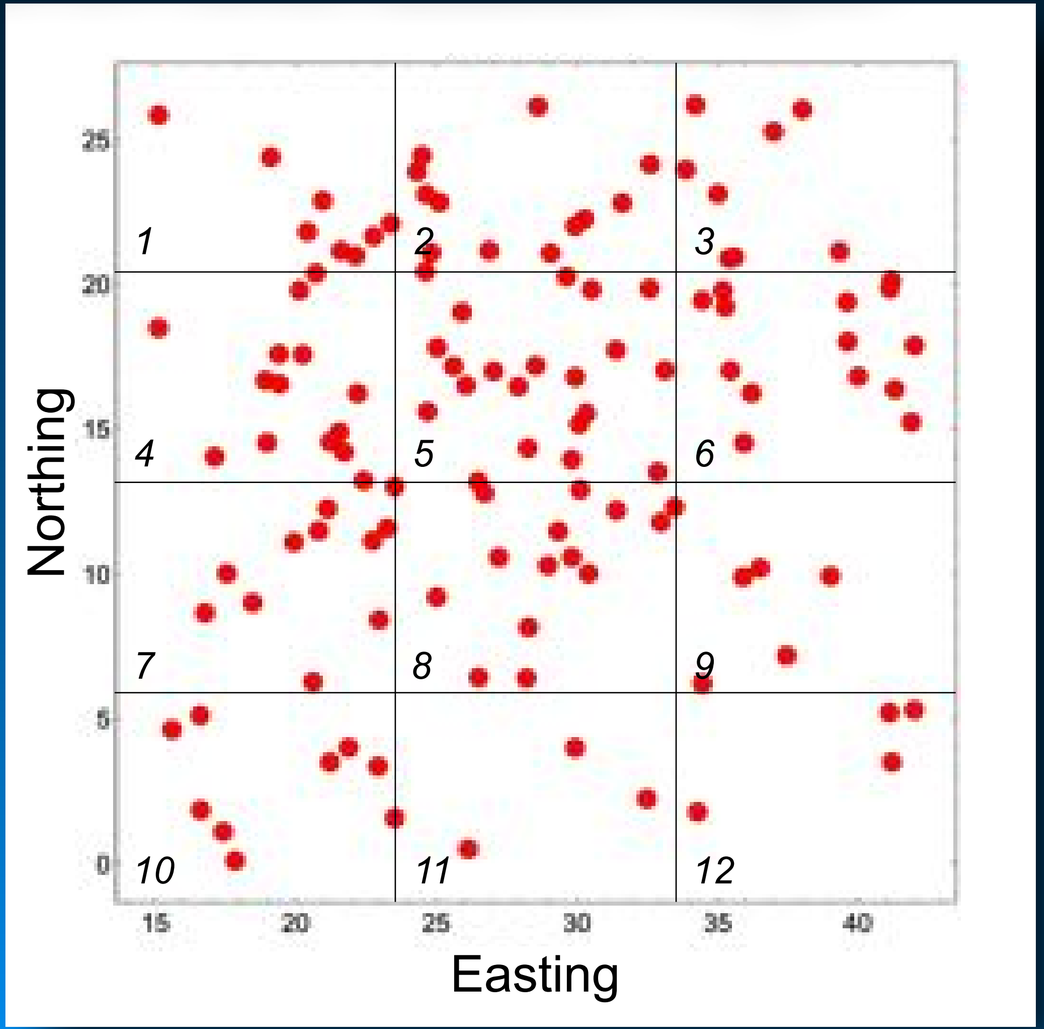


Distribution of Points

Test for Uniform Distribution of Points: Quadrat Analysis

The Chi-Squared (χ^2) Test

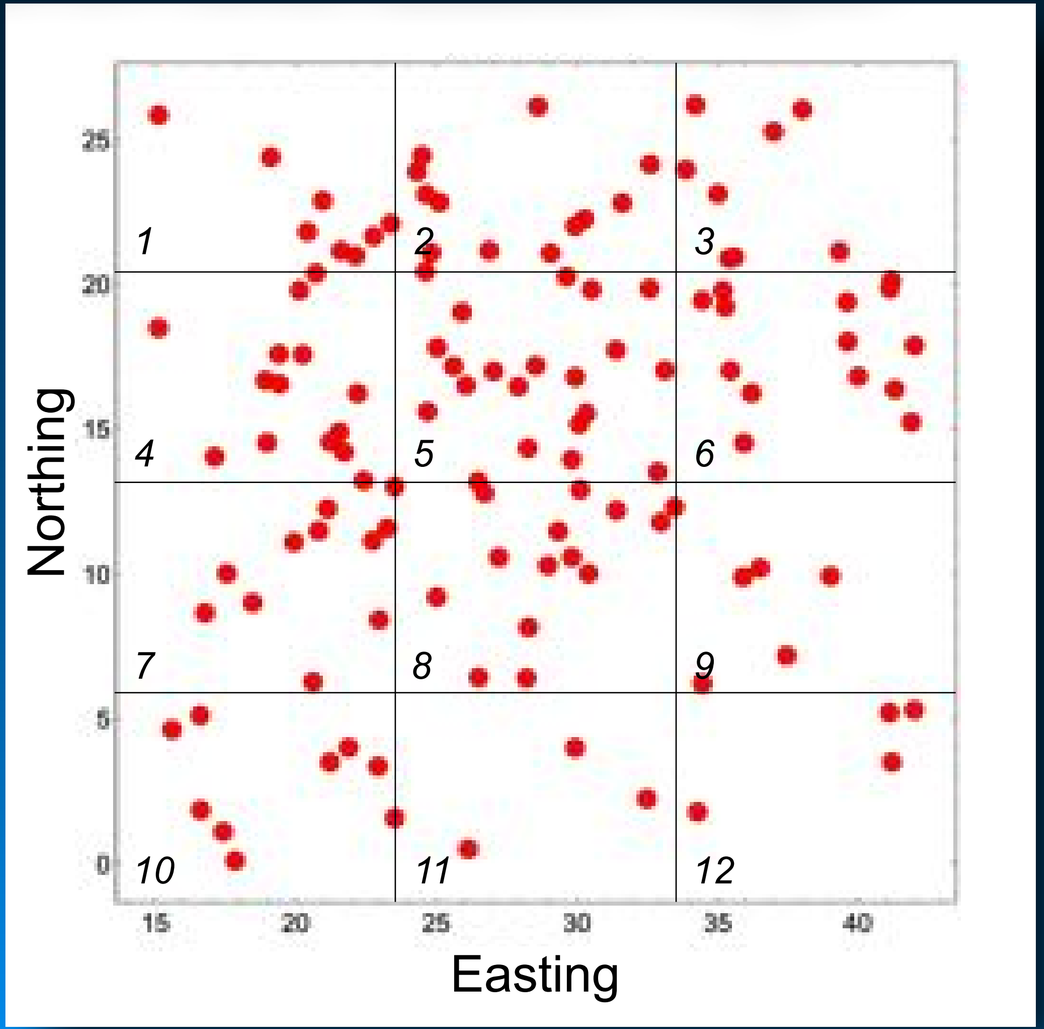
Statistical test to determine if there is a significant difference between observed and expected frequencies in categorical data. The χ^2 statistic can also be used to test for variable independence and whether the data fits a specific distribution (goodness-of-fit test).



Distribution of Points

Test for Uniform Distribution of Points: Quadrat Analysis

Cell	Obs.	Expect.	$(O-E)^2 / E$
1	8	10.25	0.49
2	13	10.25	0.74
3	8	10.25	0.49
4	14	10.25	1.37
5	20	10.25	9.27
6	14	10.25	1.37
7	11	10.25	0.05
8	14	10.25	1.37
9	5	10.25	2.69
10	9	10.25	0.15
11	3	10.25	5.13
12	4	10.25	3.81
Σ	123	123	26.95



Distribution of Points

Test for Uniform Distribution of Points: Quadrat Analysis

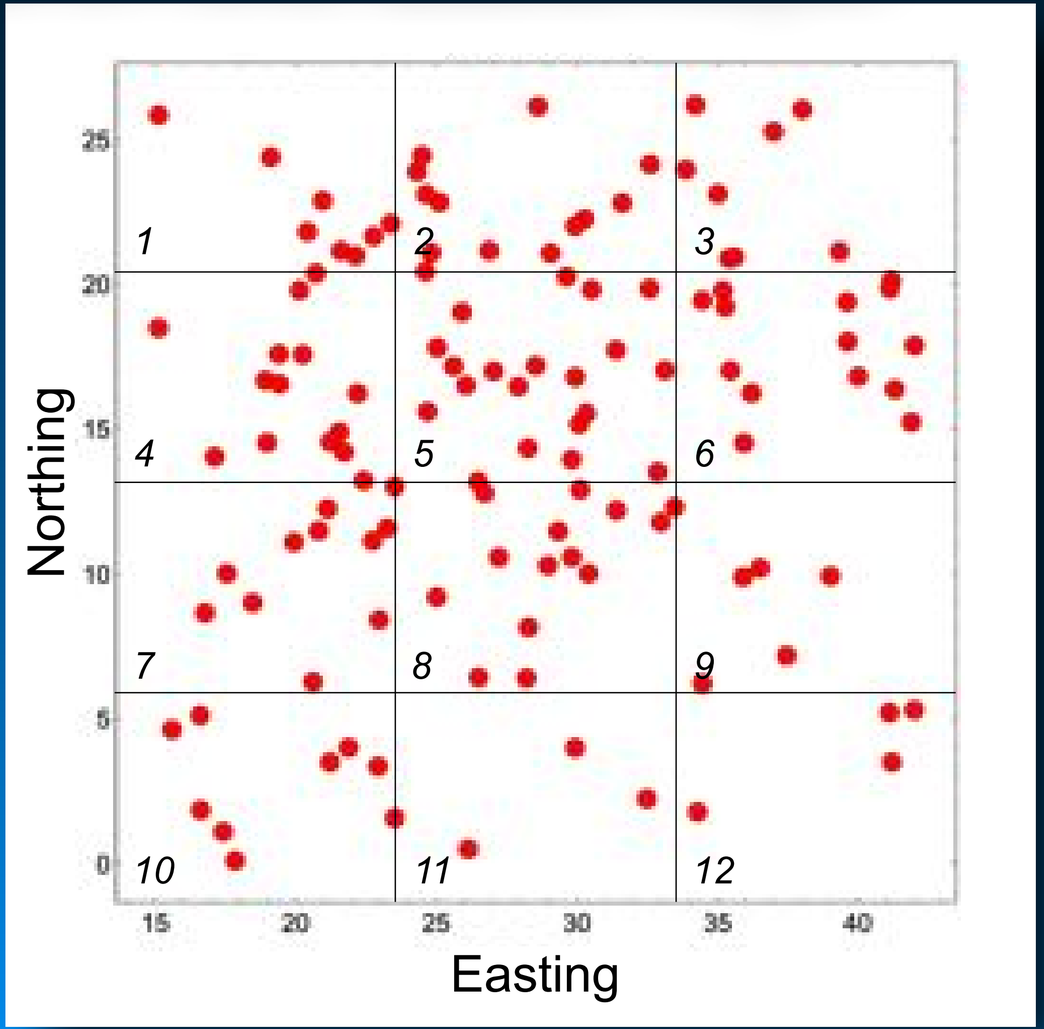
The Chi-Squared (χ^2) Test

$$\chi^2 = \sum_{i=1}^{n} \frac{(Obs. - Exp.)^2}{Exp.}$$

$$\chi^2_{\alpha=0.05, dof=10} = 18.31$$

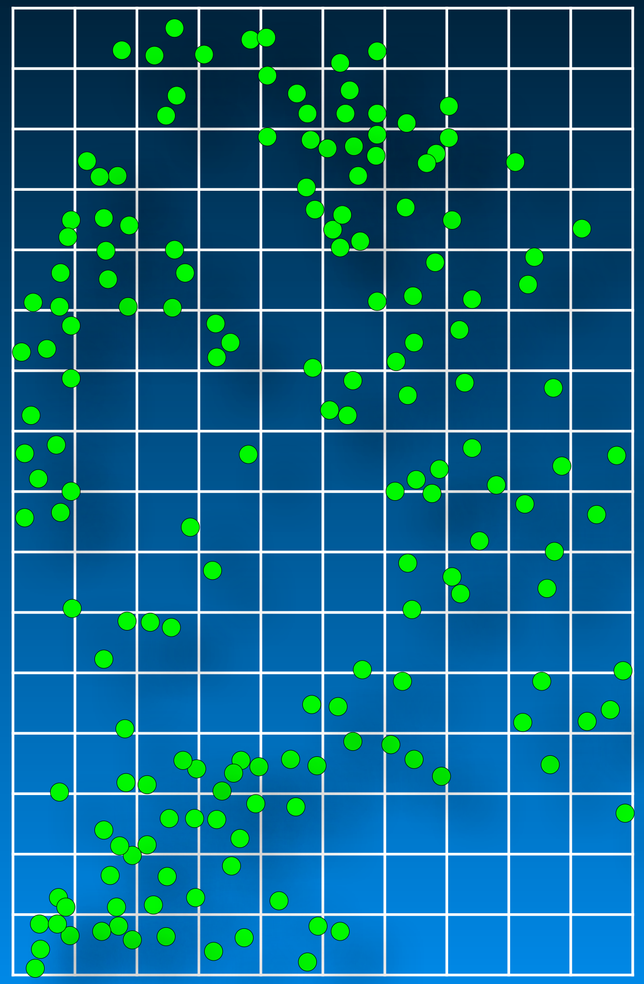
$$\chi^2_{obs.} = 26.95$$

Reject H_0



Distribution of Points

Test for a Random Distribution of Points: Quadrat Analysis

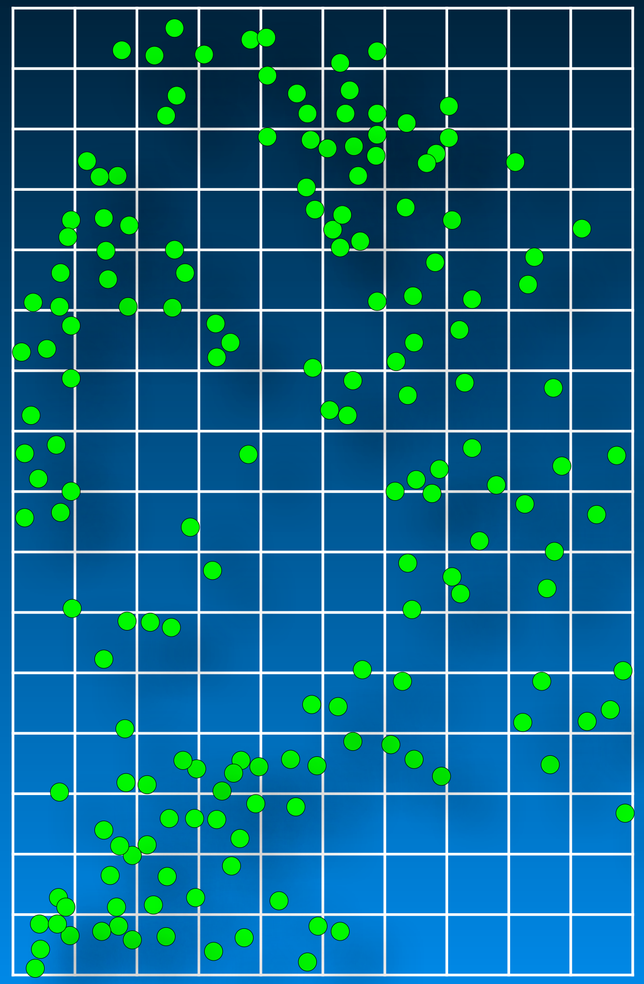


This diagram records the locations of discovery wells on the eastern shelf of the Permian Basin in West Texas. The grid represents a 10 mi^2 area and there are 168 wells in total.

Does the occurrence of oil wells in this area exhibit a random distribution?

Distribution of Points

Test for a Random Distribution of Points: Quadrat Analysis



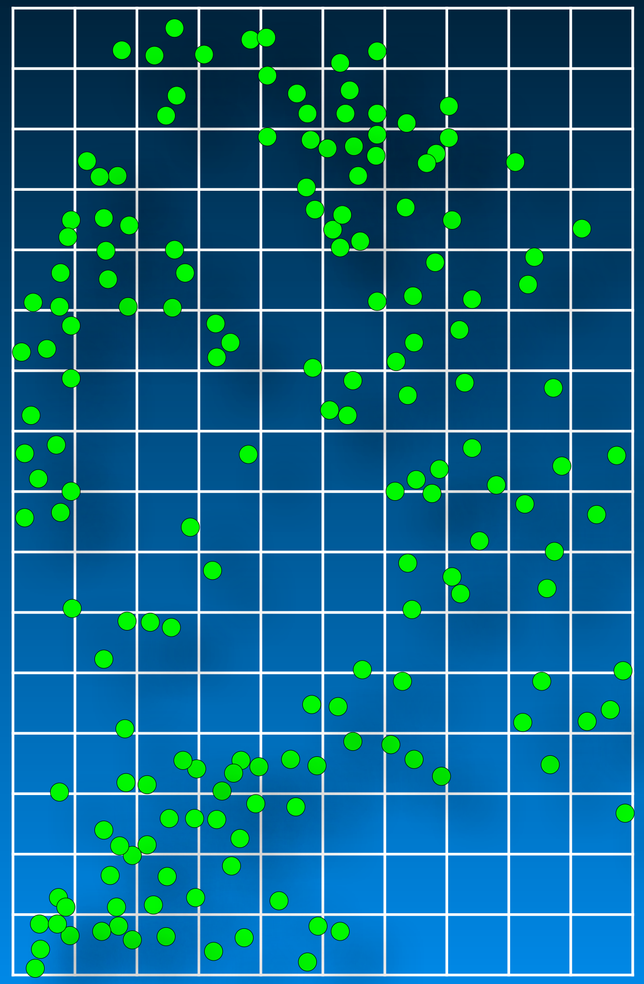
Binomial Probability Distribution of Well Occurrences per Quadrat

$$p_k = e^{-\frac{n}{T}} \cdot \left(\frac{\left(\frac{n}{T}\right)^k}{k!}\right)$$

where: $n =$ *number of observations*
$T =$ *no. of quadrats*
$k =$ *no observations per quadrat*

Distribution of Points

Test for a Random Distribution of Points: Quadrat Analysis

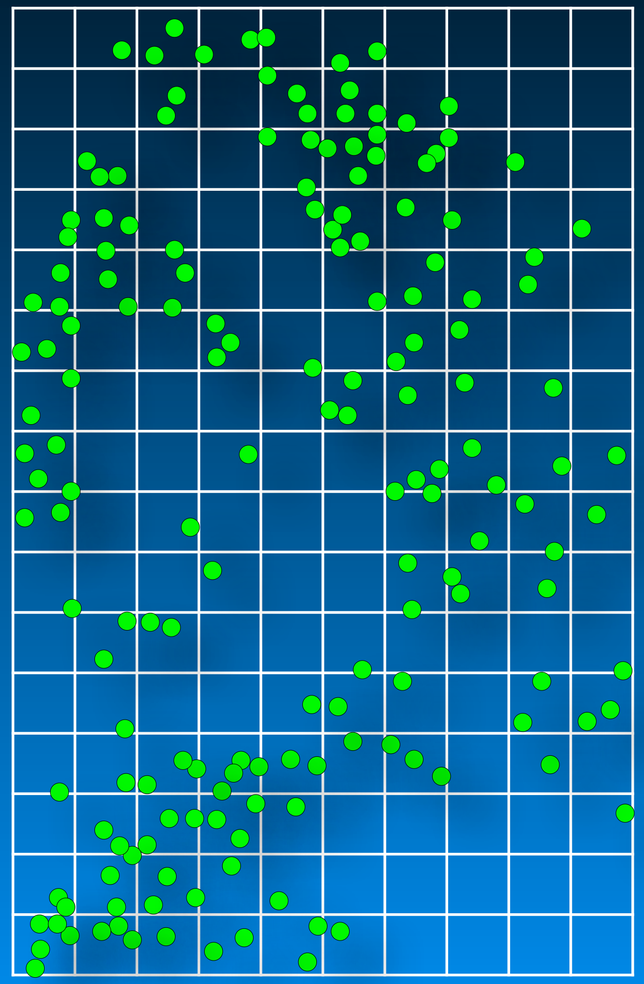


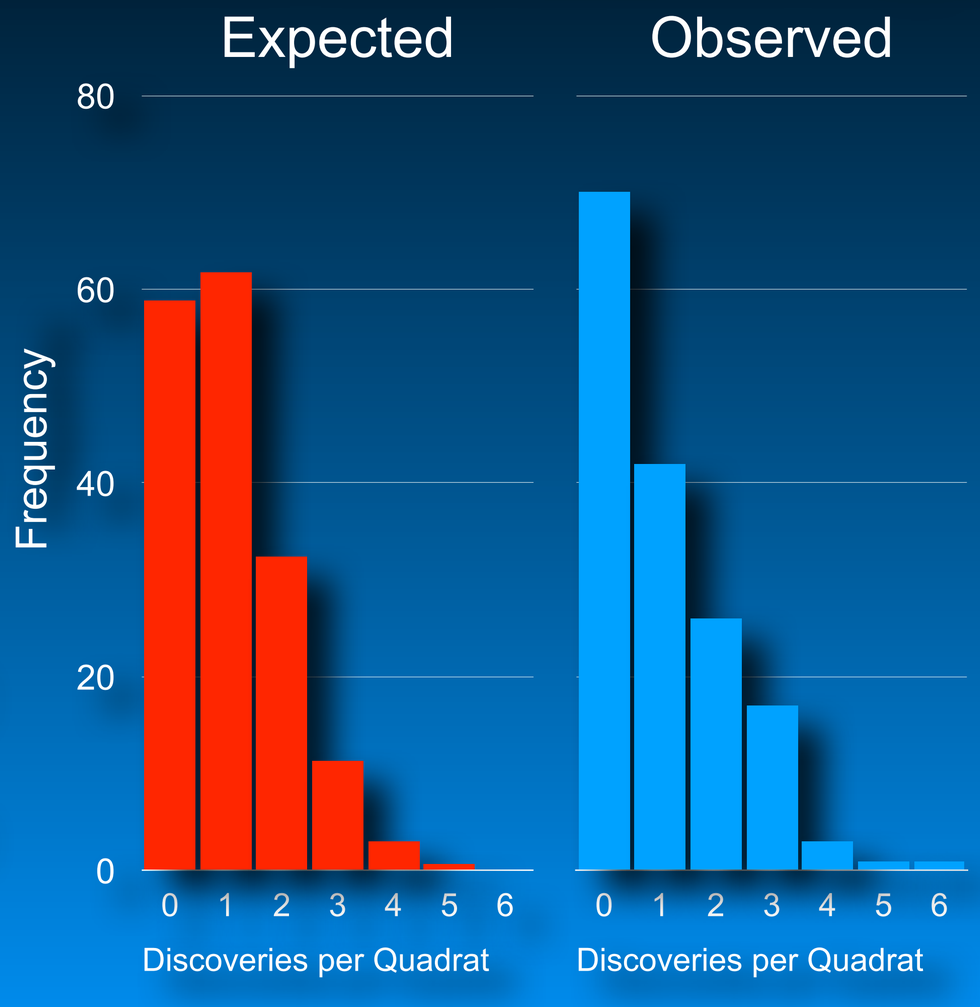
$$\chi^2 = \sum_{i=1}^{n} \frac{(Obs. - Exp.)^2}{Exp.}$$

Cell	Binomial Prob.	Expect.	Obs.	χ^2
0	0.3499	55.99	70	3.51
1	0.3674	58.79	42	4.79
2	0.1929	30.86	26	0.77
3	0.0675	10.80	17	3.56
4-6	0.0221	3.54	5	0.61
Σ	0.9999	159.98	160.00	13.23

Distribution of Points

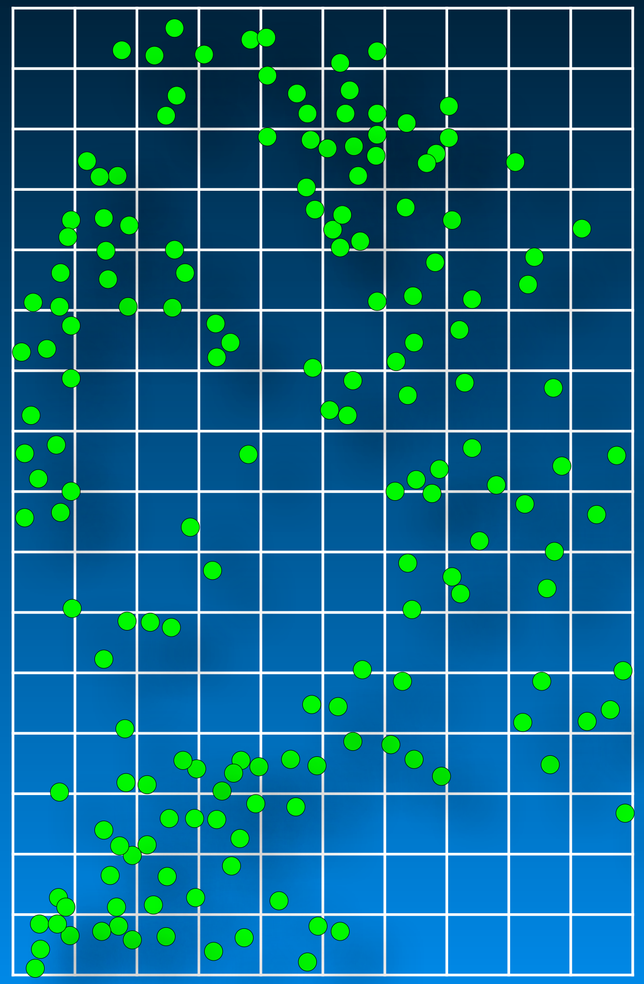
Test for a Random Distribution of Points: Quadrat Analysis





Distribution of Points

Test for a Random Distribution of Points: Quadrat Analysis



$$\chi^2 = \sum_{i=1}^{n} \frac{(Obs. - Exp.)^2}{Exp.}$$

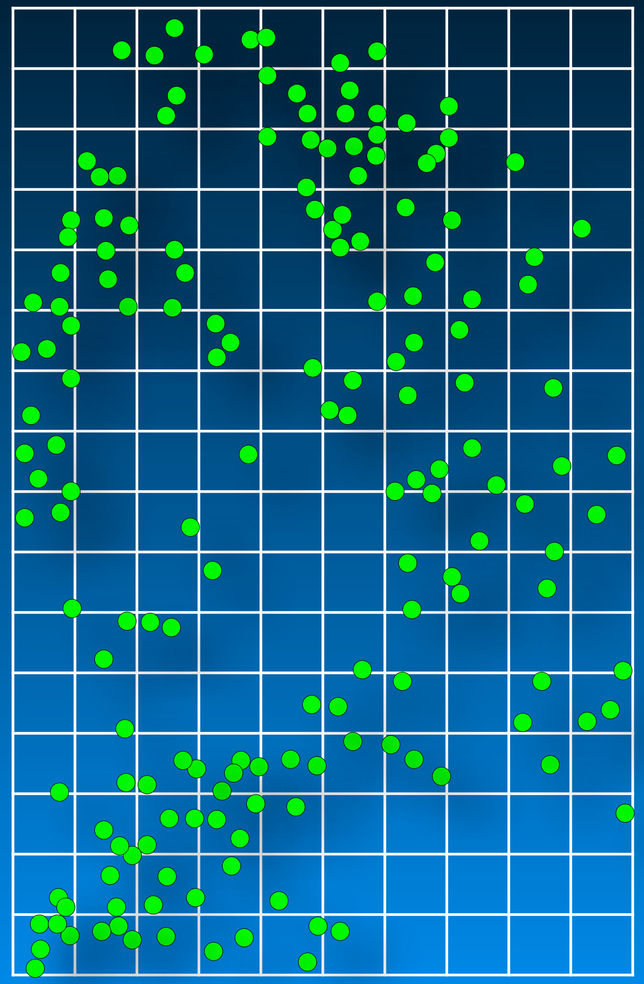
Cell	Binomial Prob.	Expect.	Obs.	x^2
0	0.3499	55.99	70	3.51
1	0.3674	58.79	42	4.79
2	0.1929	30.86	26	0.77
3	0.0675	10.80	17	3.56
4-6	0.0221	3.54	5	0.61
Σ	0.9999	159.98	160.00	13.23

$$\chi^2_{\alpha=0.05,3} = 7.81$$

$$Reject\ H_0$$

Distribution of Points

Test for a Clustered Distribution of Points: Quadrat Analysis

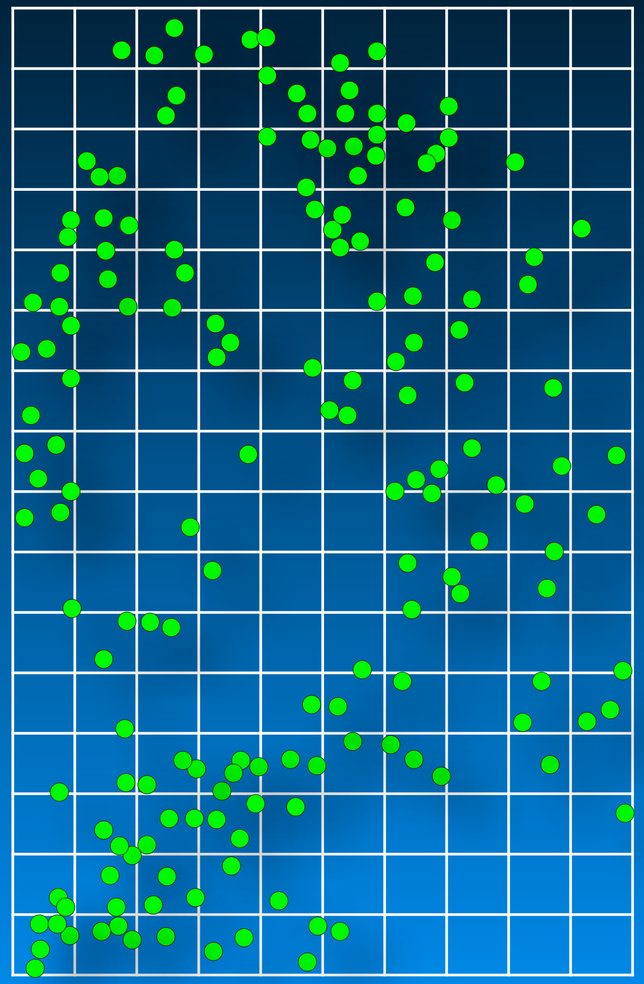


This diagram records the locations of discovery wells on the eastern shelf of the Permian Basin in West Texas. The grid represents a 10 mi^2 area and there are 168 wells in total.

Does the occurrence of oil wells in this area exhibit a clustered distribution?

Distribution of Points

Test for a Clustered Distribution of Points: Quadrat Analysis



Binomial Probability Distribution
of Well Clusters per Quadrat

$$P(0) = \frac{1}{(1+p)^k}$$

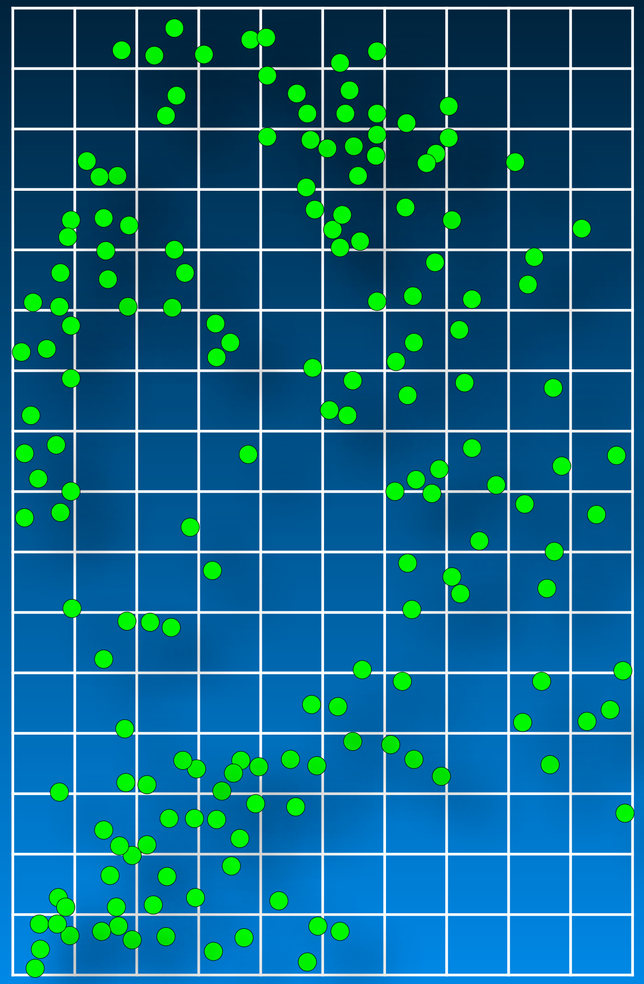
$$P(r) = \frac{(k+r-1)\cdot(\frac{p}{1+p})}{r}\cdot P(r-1)$$

Where: k = a clustering parameter scaled by the variance in discoveries across quadrats.

p = the scaled cluster probability.

Distribution of Points

Test for a Clustered Distribution of Points: Quadrat Analysis



Binomial Probability Distribution
of Well Clusters per Quadrat

$$k = \frac{(m/T)^2}{s^2 - (m/T)}$$

$$p = \frac{m/T}{k}$$

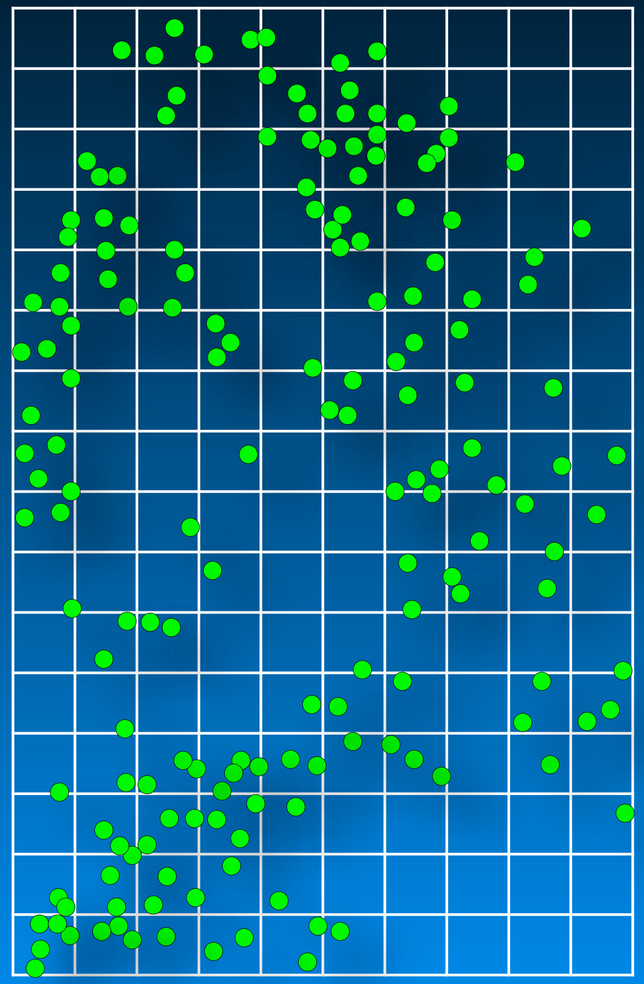
Where: m = number of discoveries.

T = number of quadrats.

s^2 = variance of discoveries across quadrats.

Distribution of Points

Test for a Clustered Distribution of Points: Quadrat Analysis

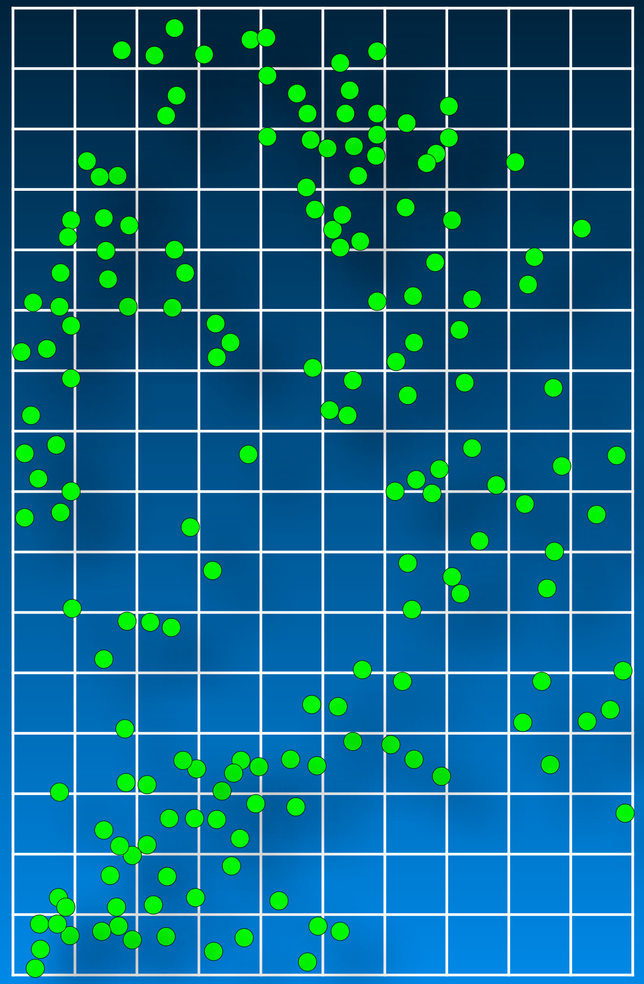


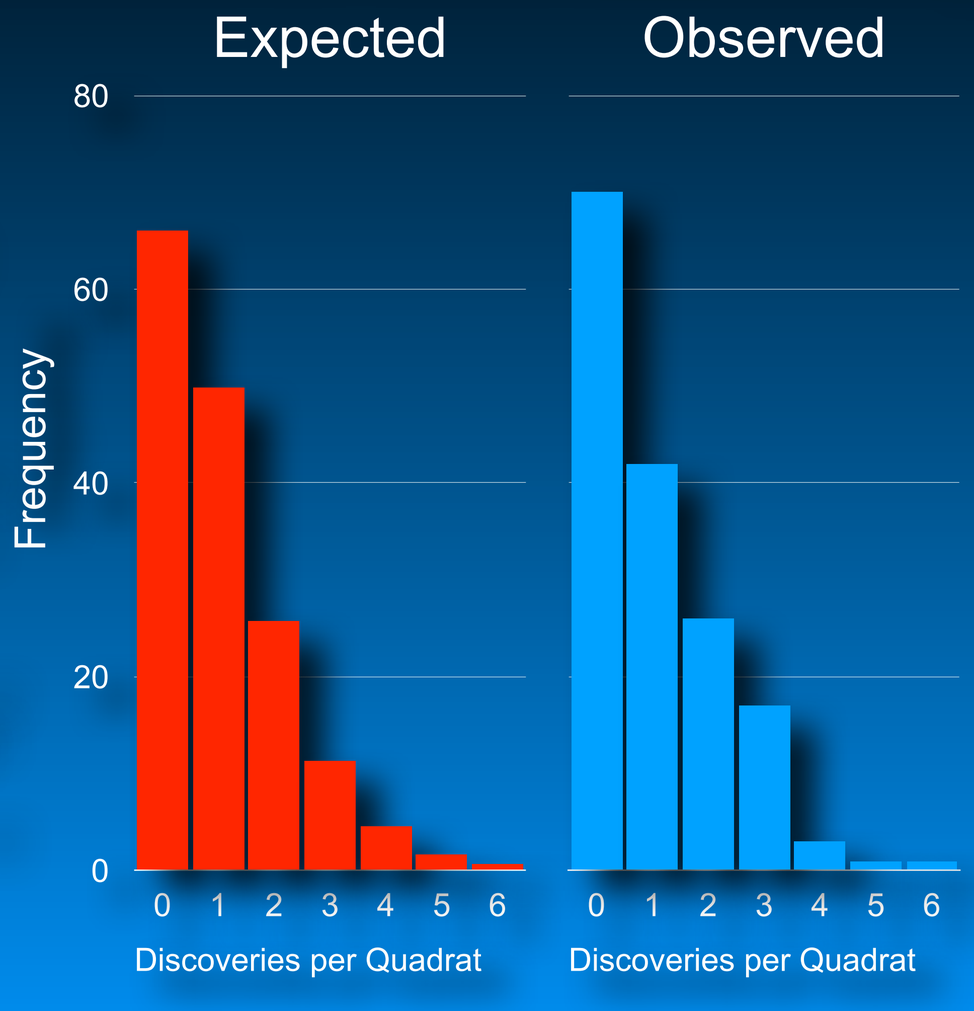
$$\chi^2 = \sum_{i=1}^{n} \frac{(Obs. - Exp.)^2}{Exp.}$$

Cell	Binomial Prob.	Expect.	Obs.	x^2
0	0.4124	65.98	70	0.24
1	0.3112	49.80	42	1.22
2	0.1611	25.78	26	0.00
3	0.0707	11.31	17	2.87
4-6	0.0426	6.81	5	0.48
Σ	0.9980	159.67	160.00	4.82

Distribution of Points

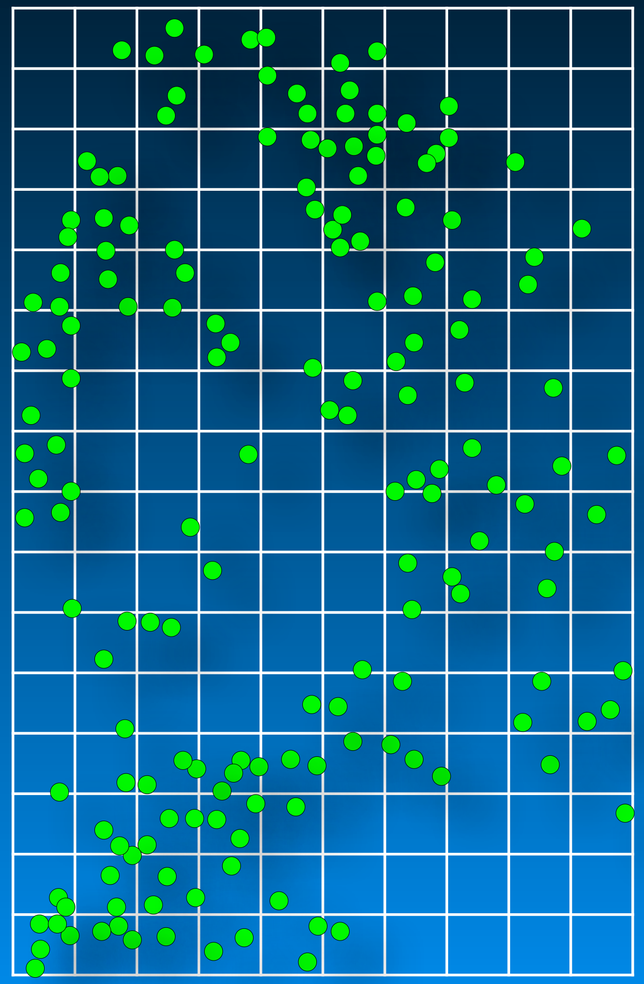
Test for a Clustered Distribution of Points: Quadrat Analysis





Distribution of Points

Test for a Clustered Distribution of Points: Quadrat Analysis



$$\chi^2 = \sum_{i=1}^{n} \frac{(Obs. - Exp.)^2}{Exp.}$$

Cell	Binomial Prob.	Expect.	Obs.	x^2
0	0.4124	65.98	70	0.24
1	0.3112	49.80	42	1.22
2	0.1611	25.78	26	0.00
3	0.0707	11.31	17	2.87
4-6	0.0426	6.81	5	0.48
Σ	0.9980	159.67	160.00	4.82

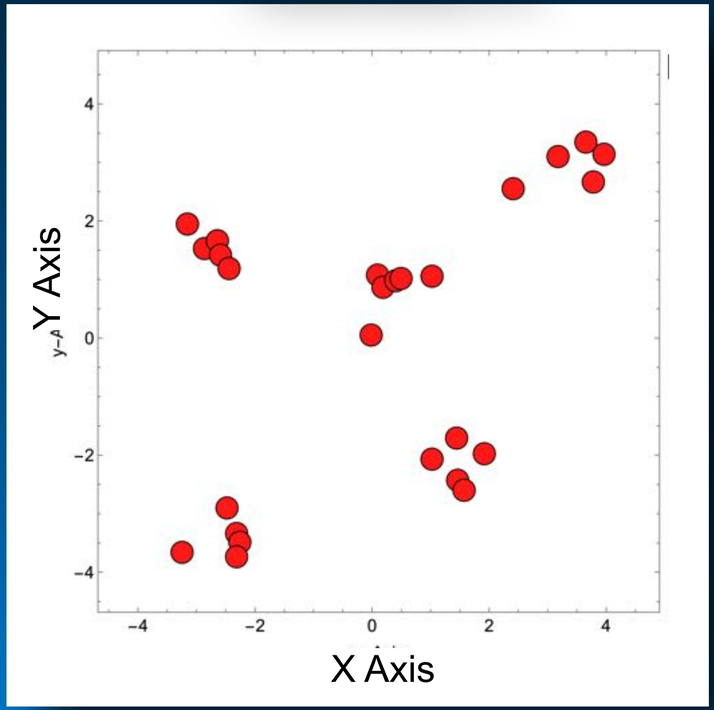
$$\chi^2_{\alpha=0.05,3} = 7.81$$

Accept H_0

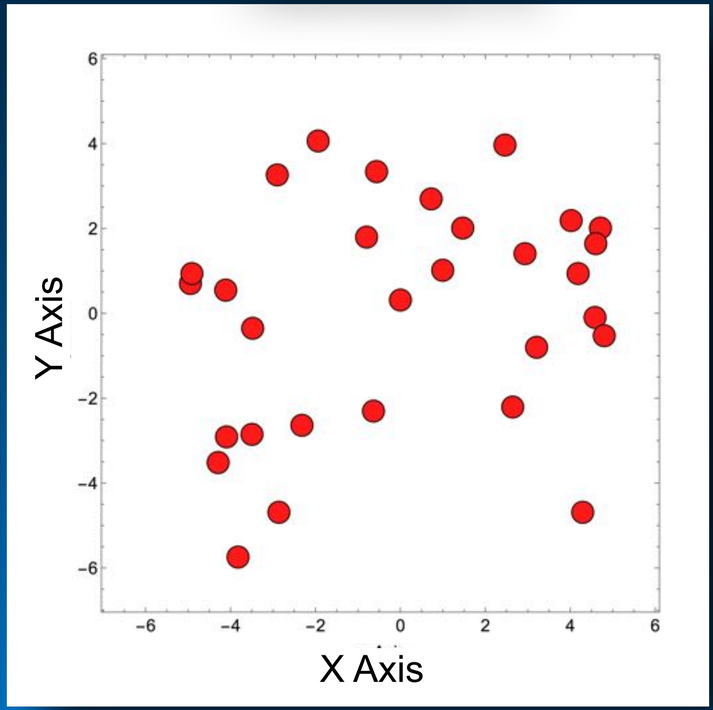
Distribution of Points

Test for a Clustered Distribution of Points: Nearest-Neighbor Analysis

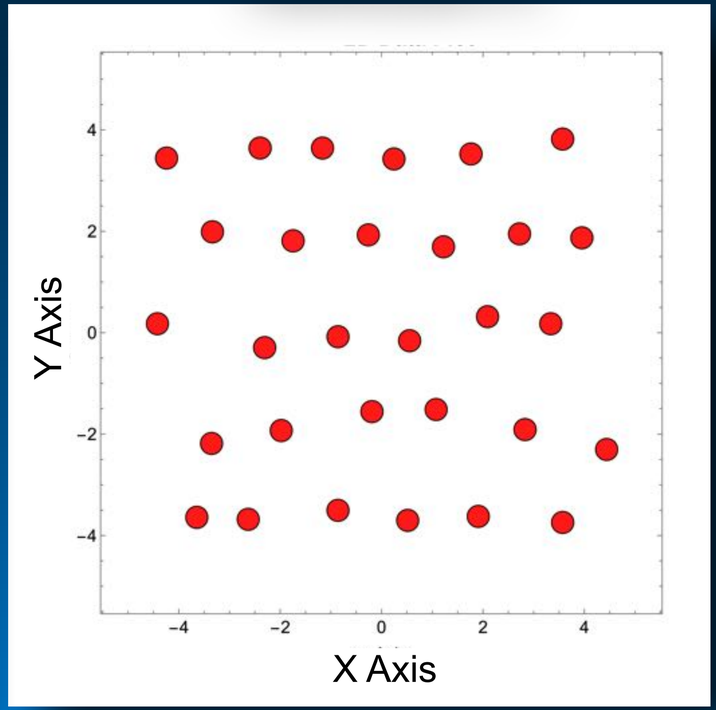
Distribution A



Distribution B



Distribution C

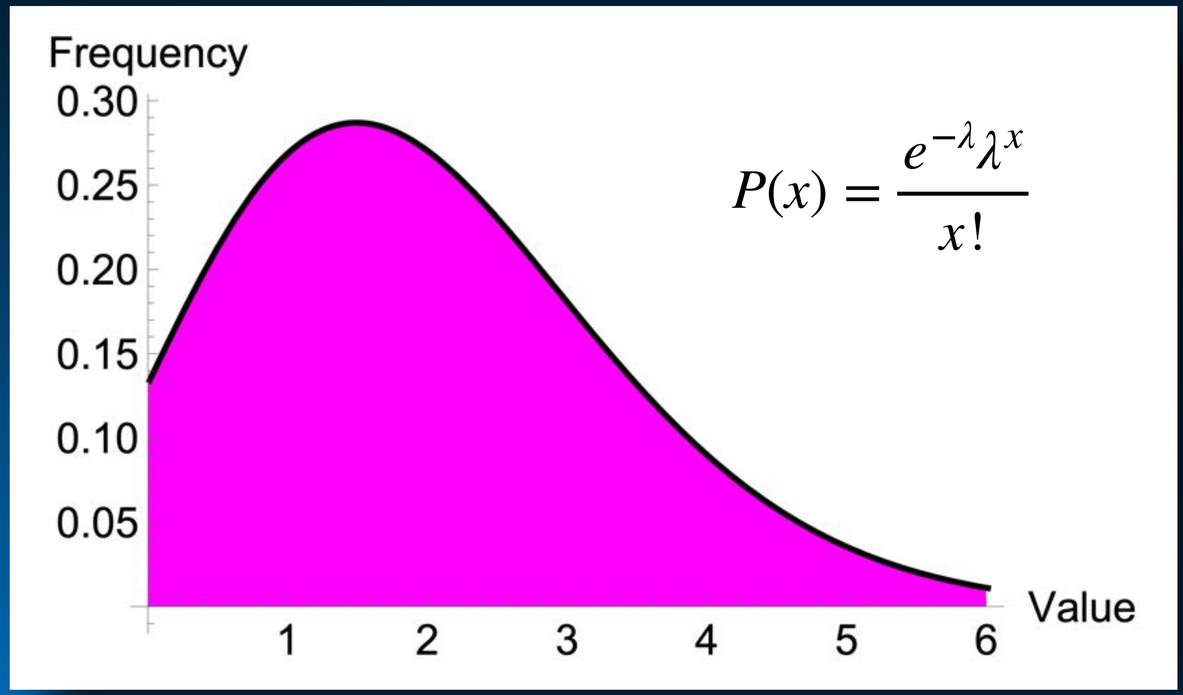


Tests also exist that forego the need to subdivide the space into quadrants. Let's compare the distributions of these three point sets and ask the question of the extent to which they exhibit clustering.

Distribution of Points

Poisson Distribution

The Poisson distribution tabulates the number of independent events (x) that occur within a fixed interval given an average rate of occurrence (λ). The Poisson distribution is useful for modelling distributions of points in space because it:

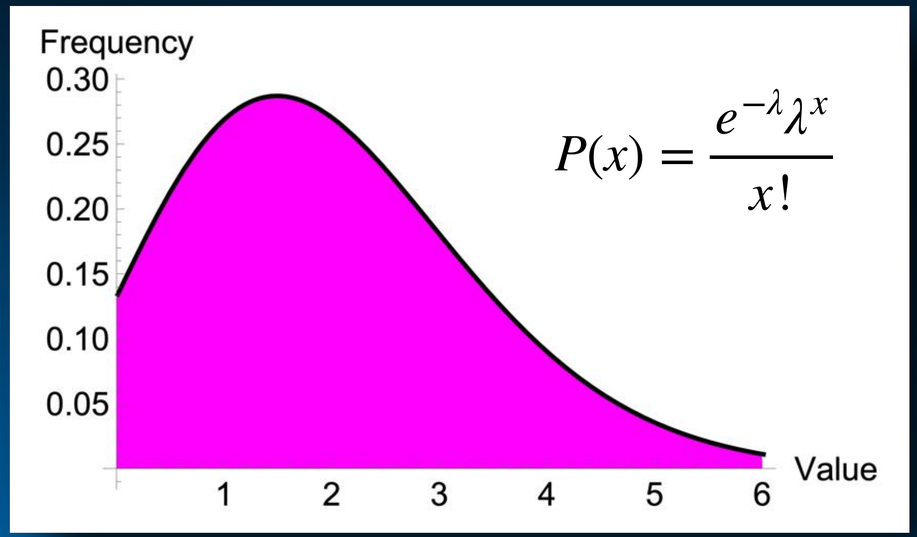


- assumes complete spatial independence (points don't attract or repel);
- places points at a constant rate (achieves homogeneity);
- works best for reproducing the occurrence of discrete events;
- models the spatial distribution of rare occurrences effectively.

Distribution of Points

Nearest-Neighbor Analysis

The Poisson distribution supports nearest-neighbor analysis because, for any given area (A) and sample size (n) it allows estimation of the mean distance between any point and its nearest neighbor for a random sample.



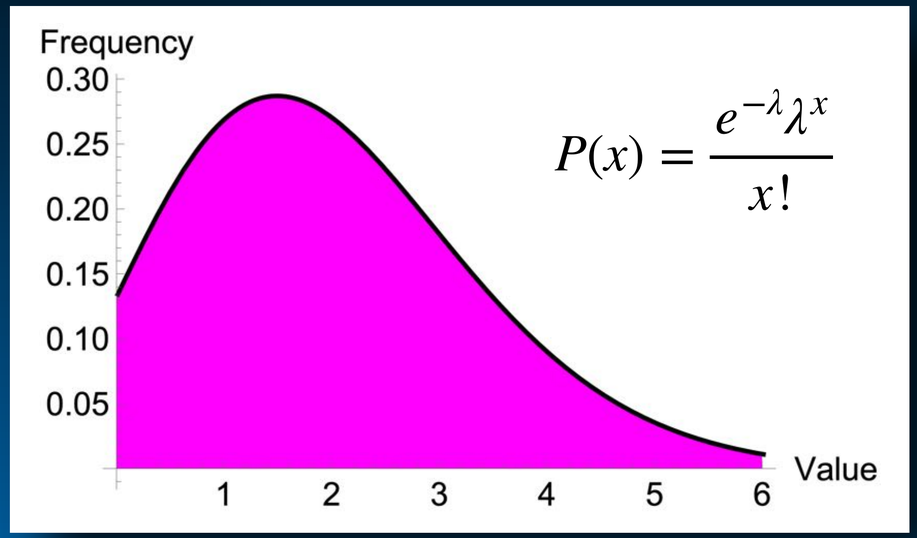
$$\bar{\mu} = \frac{1}{2}\sqrt[2]{\frac{A}{n}} \qquad R = \frac{\bar{d}}{\bar{u}}$$

Once we have this estimate we can easily calculate the mean distance to the nearest neighbor ($\bar{d}$) of all points in any distribution of points (in 2, 3, or m dimensions) and compare that value to this prediction. Clark and Evans (1954) expressed this as a ratio (R).

Distribution of Points

Nearest-Neighbor Analysis

Clark and Evans' R varies between 0.0 (all points coincide) and c. 2.15 (all points spaced regularly in a hexagonal array.



If R < 0.0 , $\bar{d} < \bar{u}$ ➛ data are clustered.

If R = 0.0 , $\bar{d} = \bar{u}$ ➛ data are distributed randomly.

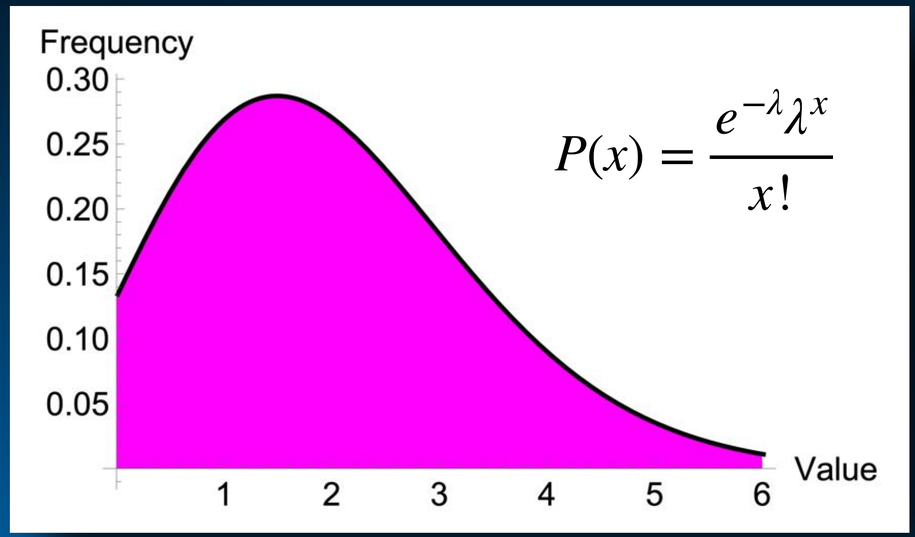
If R > 0.0 , $\bar{d} > \bar{u}$ ➛ data are dispersed.

Note, this test provides not only a means whereby we can identify when data exhibit a clustered character (as well as providing a definition of what clustered data are), it also provides a means whereby we can identify data that have a non-random dispersed character (as well as providing a definition of what dispersed data are).

Distribution of Points

Nearest-Neighbor Analysis

The Poisson distribution also allows us to turn this relation into a parametric statistical test. To do this we need an estimate of the standard error for the estimate of the mean nearest-neighbor distance (s_e).



$$s_e = \frac{\sqrt{(4-\pi)A}}{2\sqrt[2]{\pi n}} \qquad R = \frac{\bar{d}}{\bar{u}}$$

Once we have this estimate we can easily calculate the associated probability (p) value via reference to the standardized normal distribution (Z test). Note this test does not assume the data are normally distributed because (i.) $\bar{u}$ is estimated from the Poisson distribution and (ii.) $\bar{d}$ will be distributed normally via the CLT.

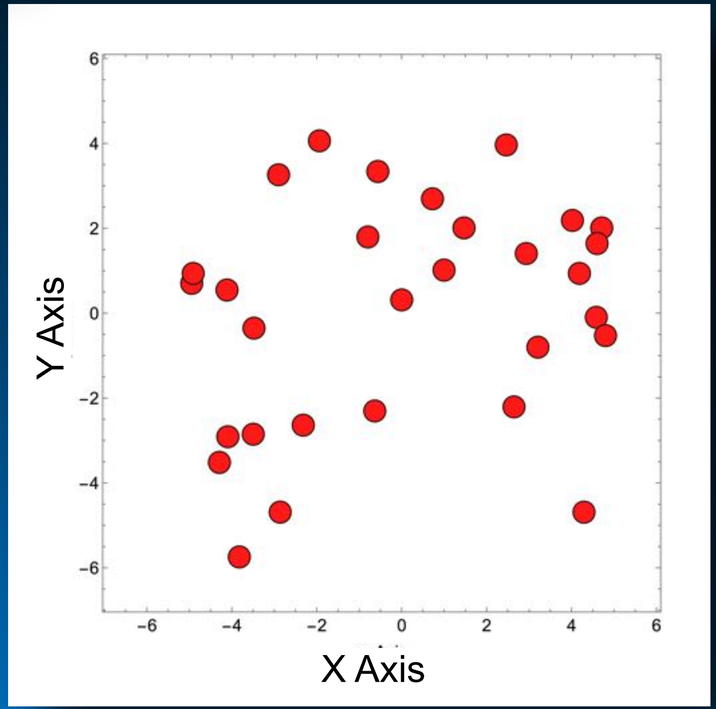
Distribution of Points

Nearest-Neighbor Analysis

One widely recognized problem with nearest-neighbor analysis is that points near the edge of the distribution have fewer neighbors than points close to the distribution's center. There are serval potential solutions to this problem, the most commonly implemented of which is a set of correction factors for the estimate of the mean nearest-neighbor distance and its variance (see Donnelly, 1978) based on extensive simulations.

$$\bar{\delta} \approx \frac{1}{2}\sqrt[2]{\frac{A}{n}} + (0.514 + \frac{0.412}{\sqrt[2]{n}})\frac{P}{n}$$

$$s_\delta^2 \approx 0.070\frac{A}{n^2} + 0.035P\frac{\sqrt[2]{A}}{n^{\frac{5}{2}}}$$

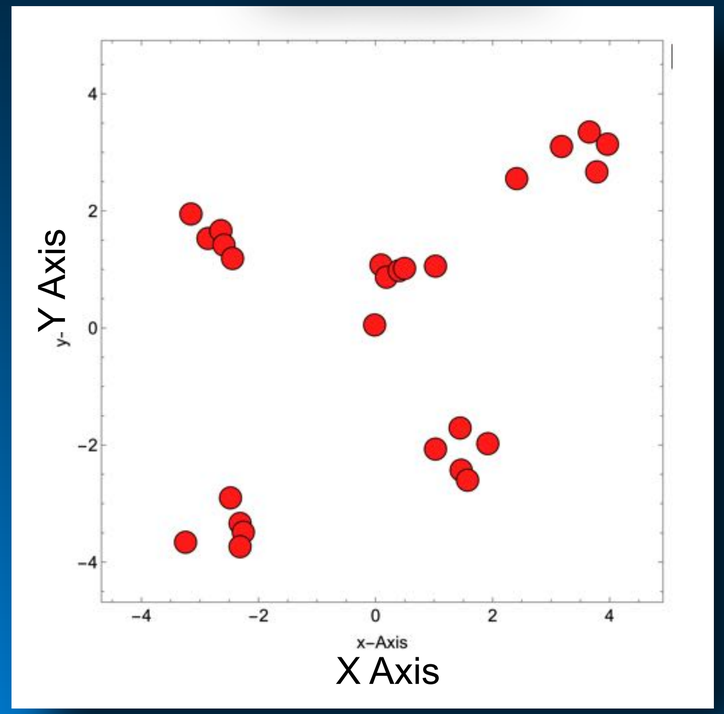


Note: in both these equations P is the perimeter of the boundary used to estimate A.

Distribution of Points

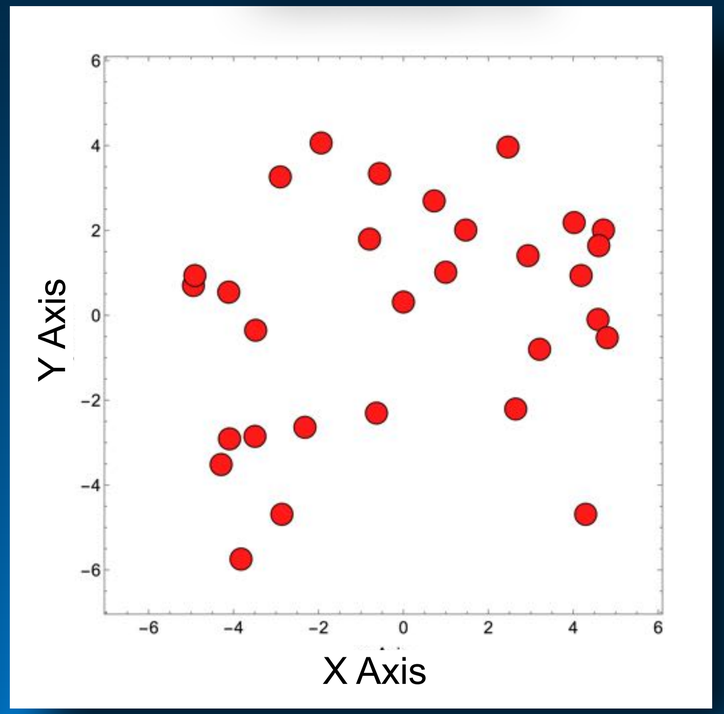
Test for a Clustered Distribution of Points: Nearest-Neighbor Analysis

Distribution A



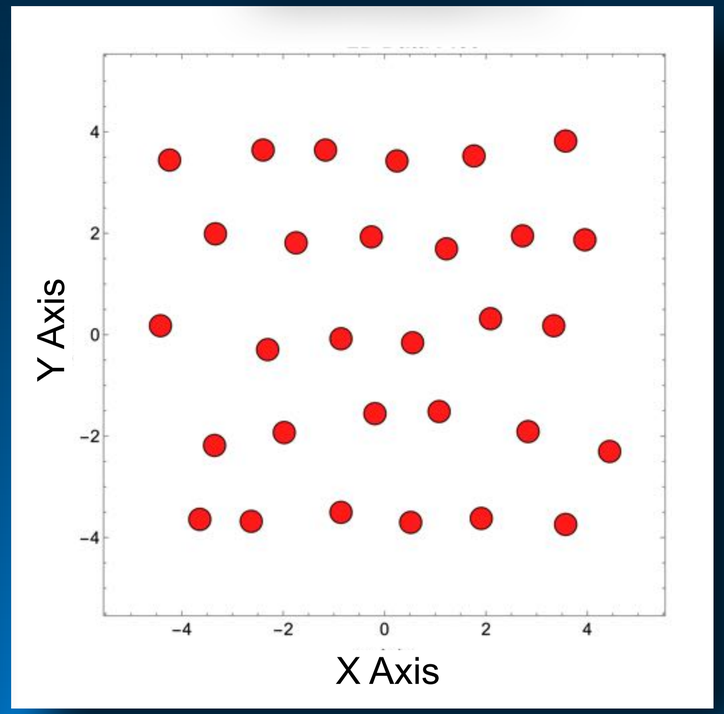
$\bar{u}$ = 0.574 R = 0.691
$\bar{d}$ = 0.397 z = 3.017
a = 0.0013

Distribution B



$\bar{u}$ = 0.811 R = 1.341
$\bar{d}$ = 1.088 z = 3.570
a = 0.0001

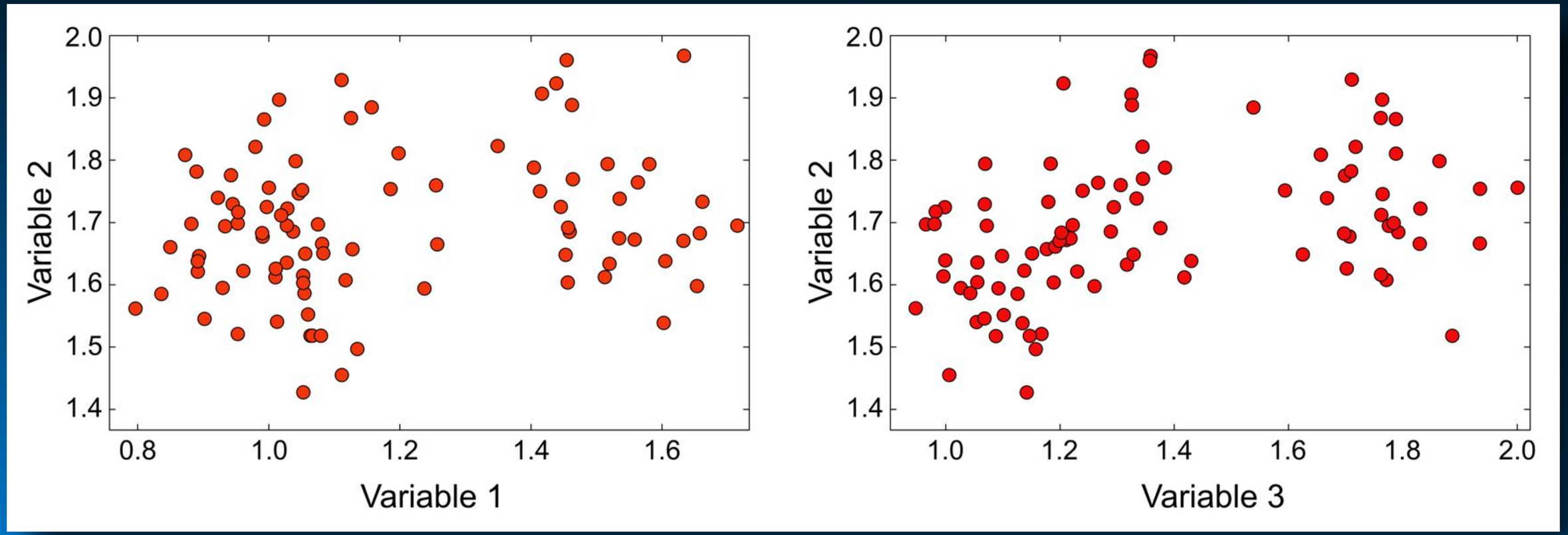
Distribution C



$\bar{u}$ = 0.712 R = 2.010
$\bar{d}$ = 1.430 z = 10.581
a = 0.000000000

Distribution of Points

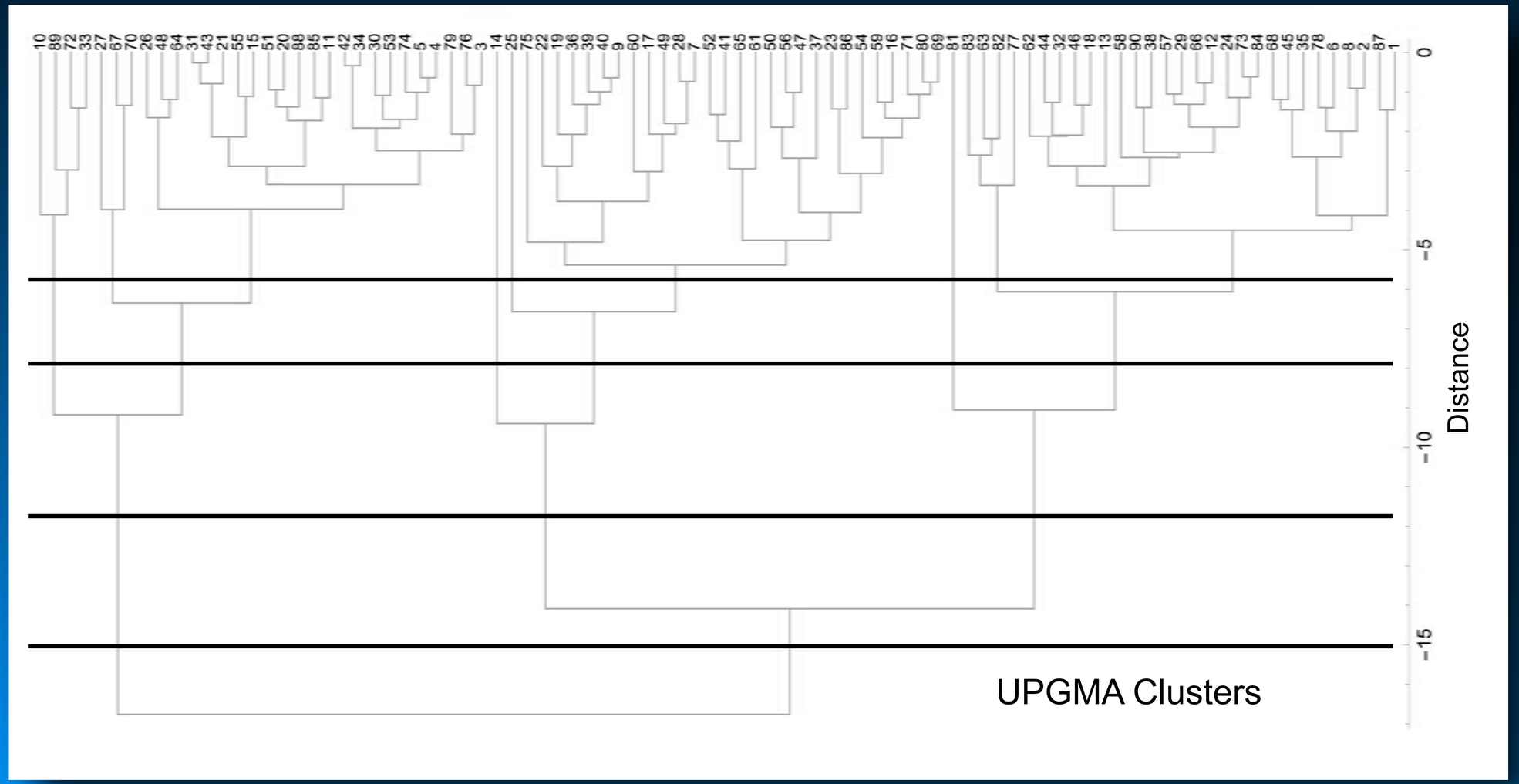
How Many Clusters are Present in a Dataset



In earth science research we are often confronted with ordination plots that look like this and asked to determine how many clusters of points are present in such data. For low-dimensional data this question can usually be addressed qualitatively via plotting. But with high-dimensional data this is not a practical option.

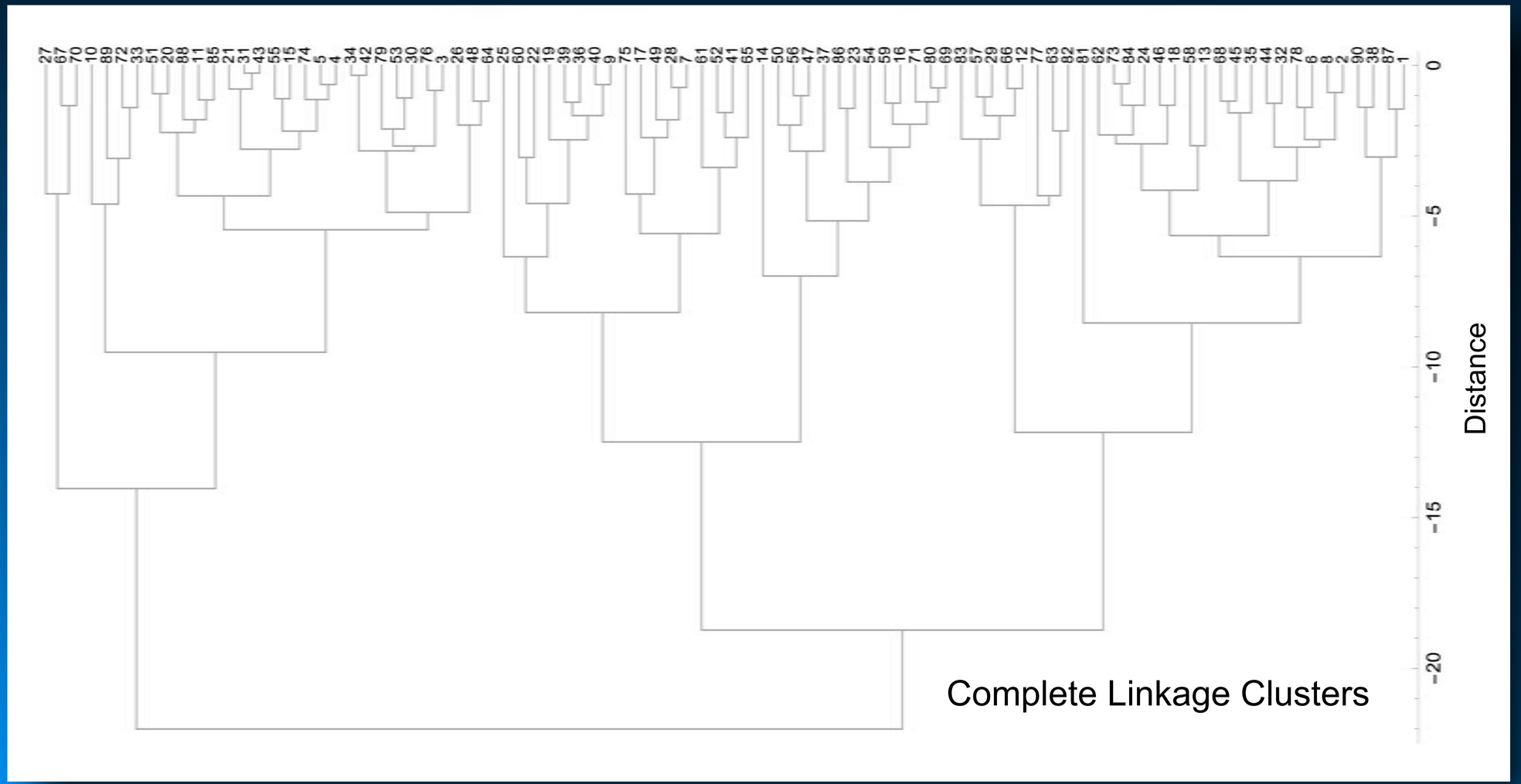
Distribution of Points

How Many Clusters are Present in a Dataset?



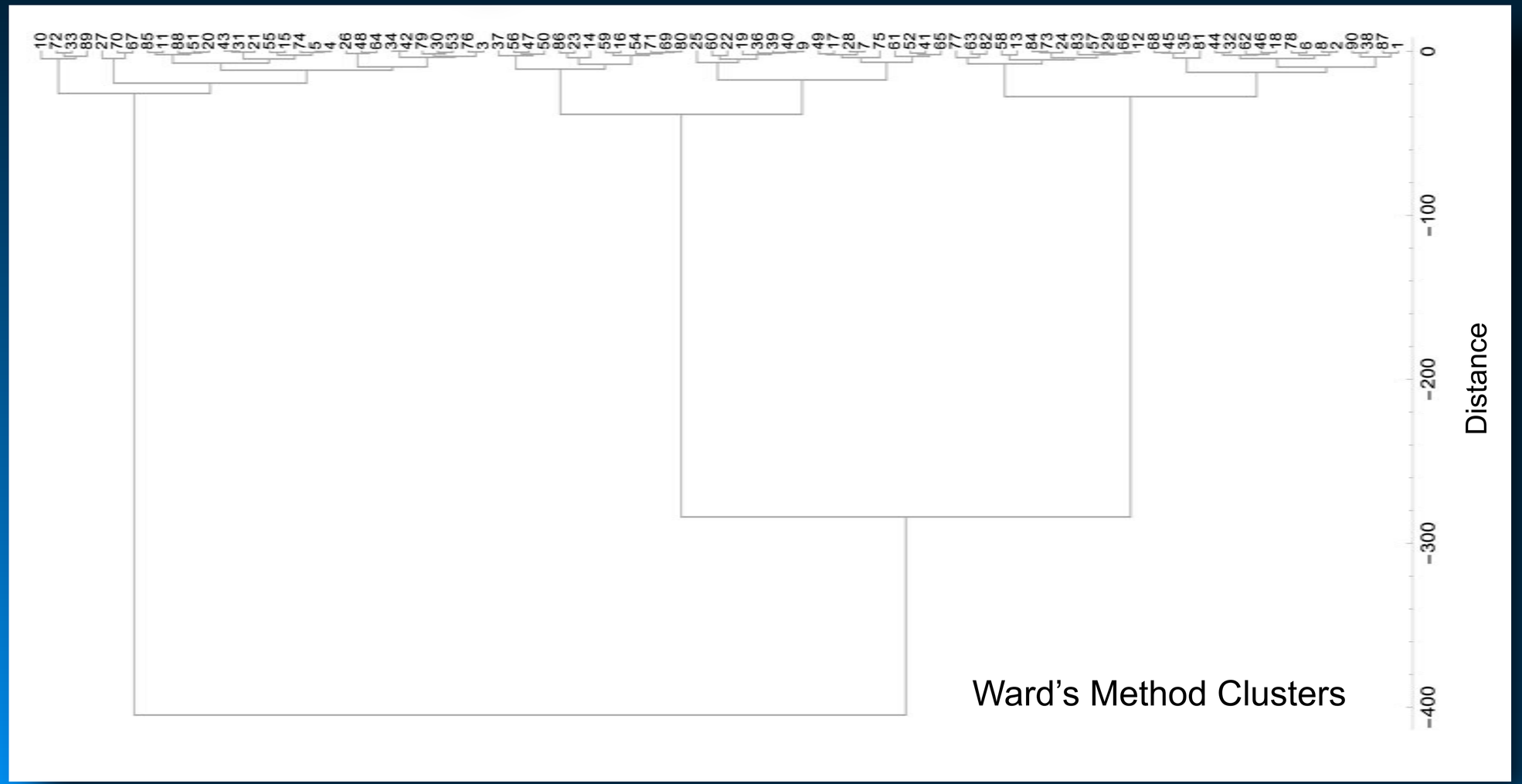
Distribution of Points

How Many Clusters are Present in a Dataset?



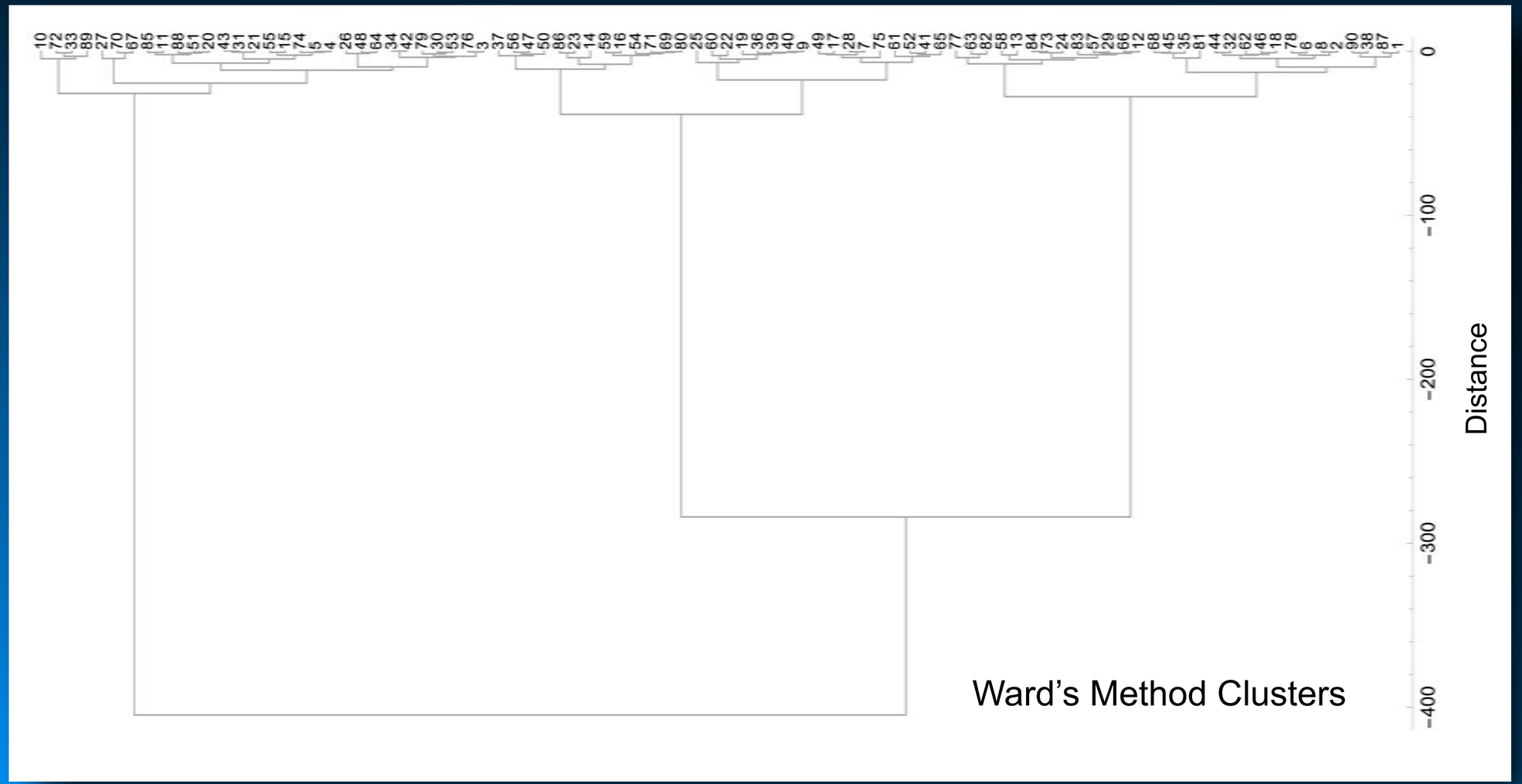
Distribution of Points

How Many Clusters are Present in a Dataset?



Distribution of Points

How Many Clusters are Present in a Dataset?



Distribution of Points

Cluster-Quality Indices

Davies-Bouldin Index

$$S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i}\left|X_j - A_i\right|^q\right)^{1/q}$$

Dispersion within each cluster (i)

$$M_{ij} = \left(\sum_{k=1}^{N}\left|A_{ki} - A_{kj}\right|^p\right)^{1/p}$$

Dispersion of cluster centroids (i vs j)

$$R_{ij} = \frac{S_i + S_j}{M_{ij}}$$

Similarity between clusters (i vs j)

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} max_{j:j\neq 1} R_{ij}$$

Average similarity

Davies & Bouldin (1979)

Distribution of Points

Cluster-Quality Indices

Rousseeuw's Silhouette Index

$$a_{i,k} = \frac{1}{T_k - 1} \sum_{j=1:j\neq i}^{T_k} d_{j,k}$$

Average distance between each point (a_{ik}) in each cluster (k) and every other point in that cluster. The choice of the distance metric (d) to use is up to the analyst and should be made based on the character of the data and purpose of the analysis.

$$b_{i,l} = min_{j\neq i} \frac{1}{T_l} \sum_{j=1}^{T_l} d_{j,l}$$

Comparison between the points (a_{ik}) in each cluster and the cluster nearest to k to which the point (a_{ik}) does not belong.

Rousseeuw (1987)

Distribution of Points

Cluster-Quality Indices

Rousseeuw's Silhouette Index

$$s_i = \begin{cases} 1 - \frac{a_i}{b_i} \text{ , if } a_i < b_i \\ 0 \text{ , if } a_i = b_i \\ \frac{a_i}{b_i} - 1 \text{ , if } a_i > b_i \end{cases}$$

A comparison is then made between the intra-cluster distances (a) and inter-cluster (b) distances.

The Silhouette Index is bounded between -1 and 1. A large positive s_i value suggests the object has been placed in the correct cluster, a low positive or negative s_i value suggests the object's assignment is equivocal. The average of the s_i values ($\bar{s}$) is an overall index for the quality of the cluster solution.

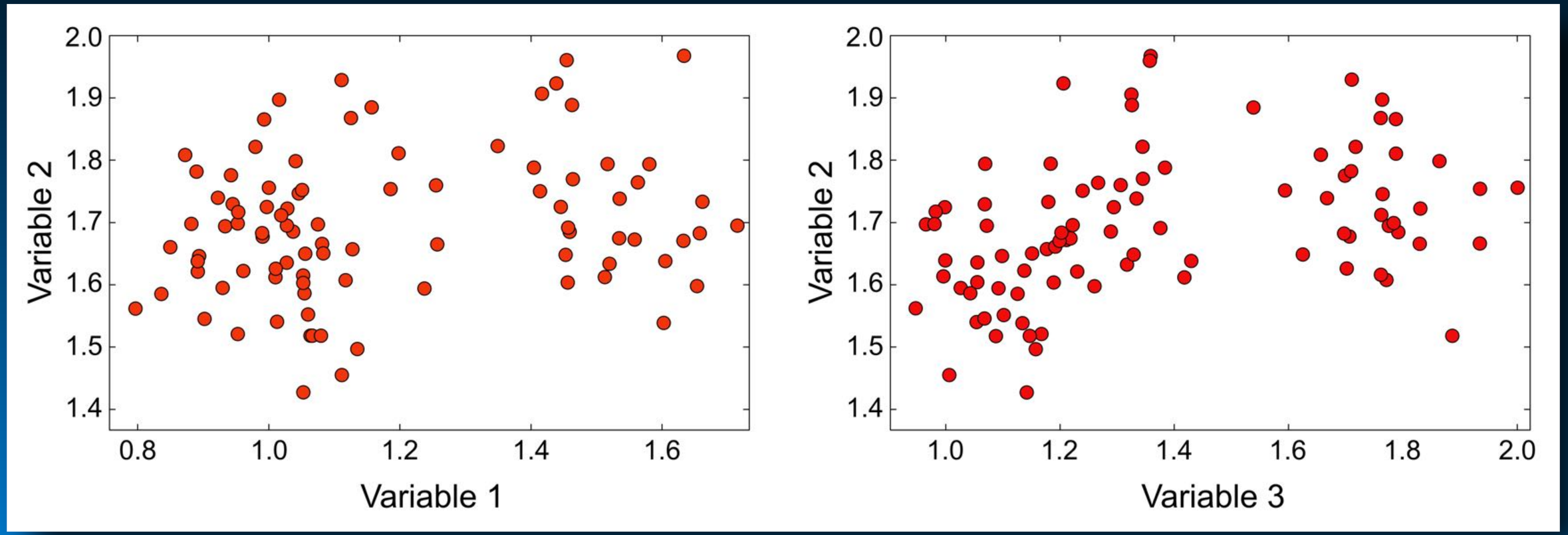
Rousseeuw (1987)

Distribution of Points

Cluster-Quality Indices

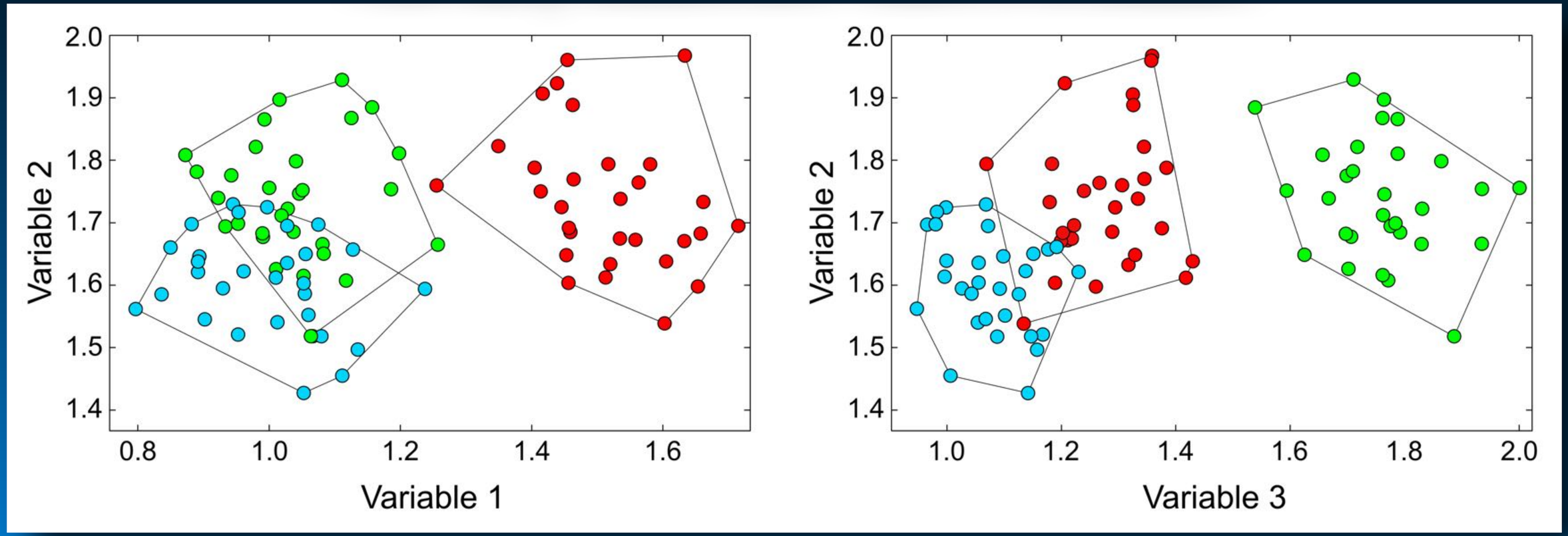
Distribution of Points

How Many Clusters are Present in a Dataset?



Distribution of Points

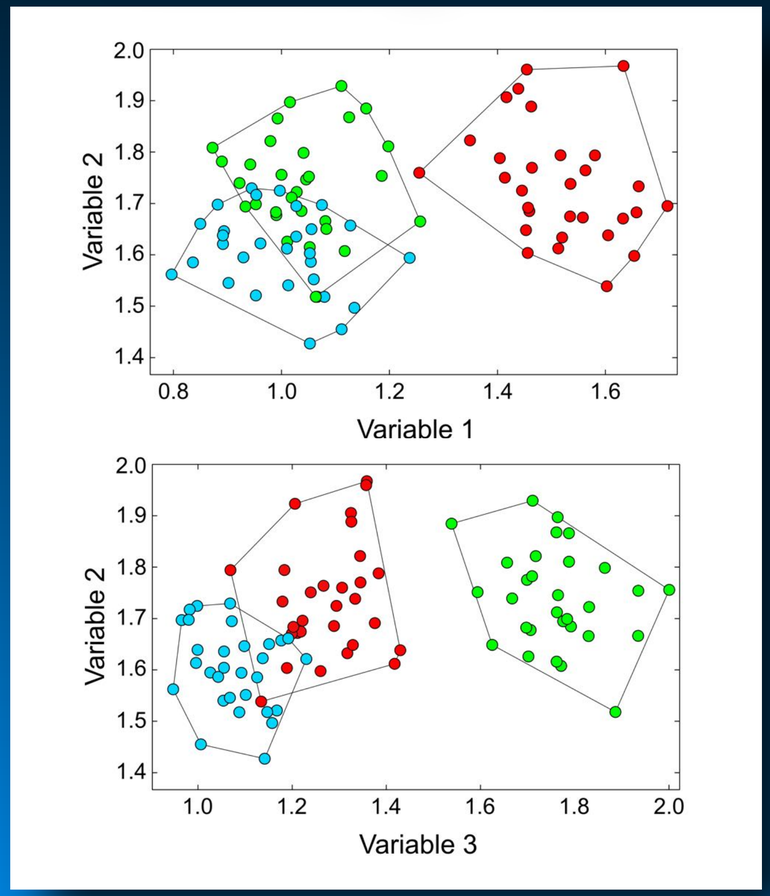
How Many Clusters are Present in a Dataset?

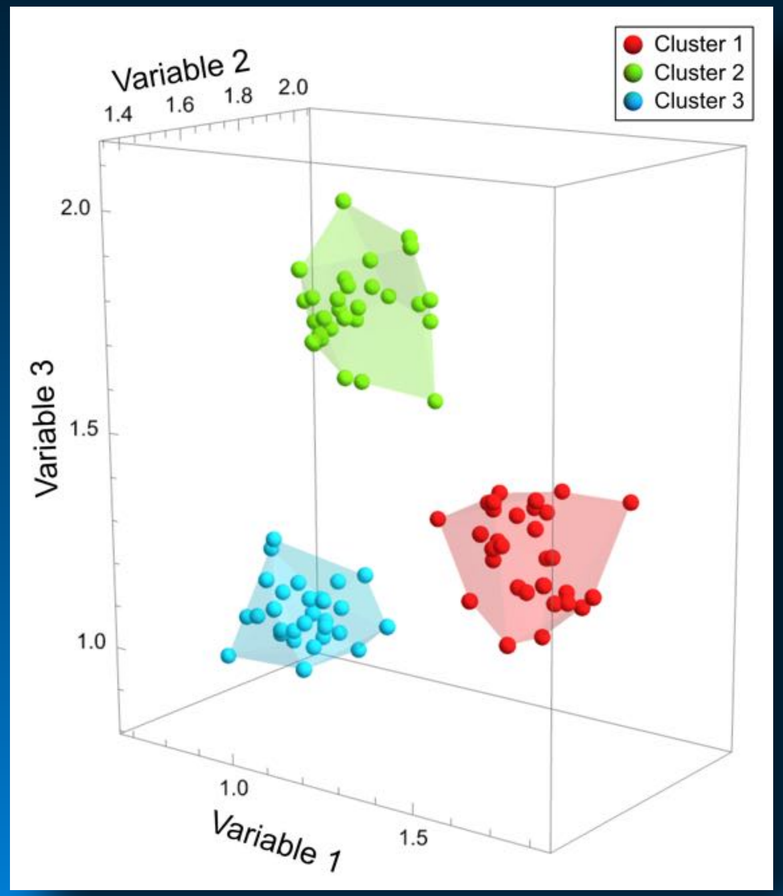


Optimal cluster solution (K = 3) for the example dataset.

Distribution of Points

How Many Clusters are Present in a Dataset?

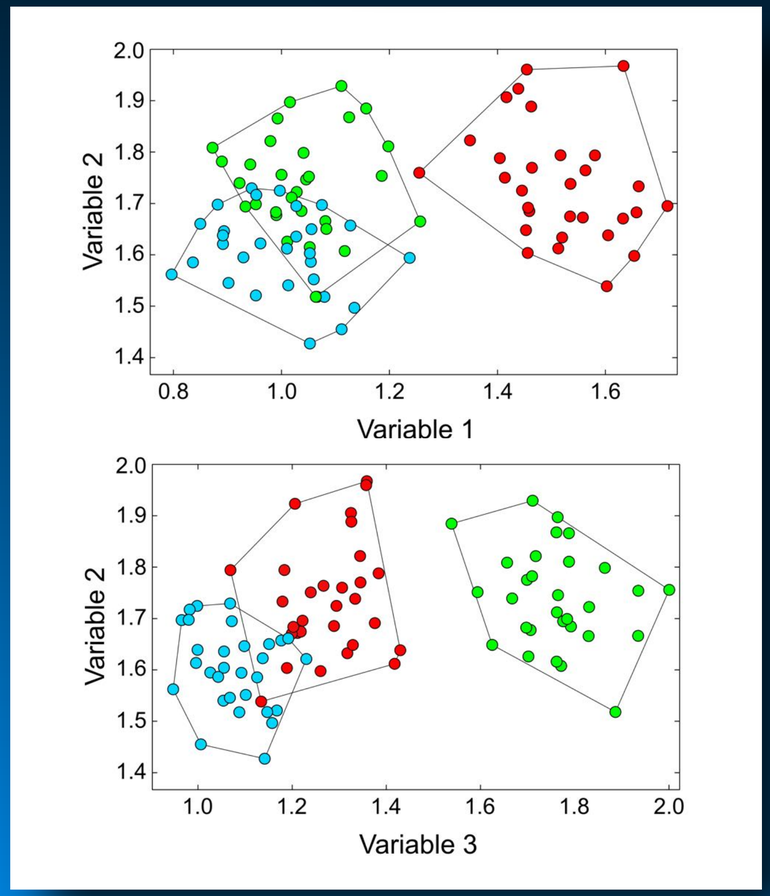


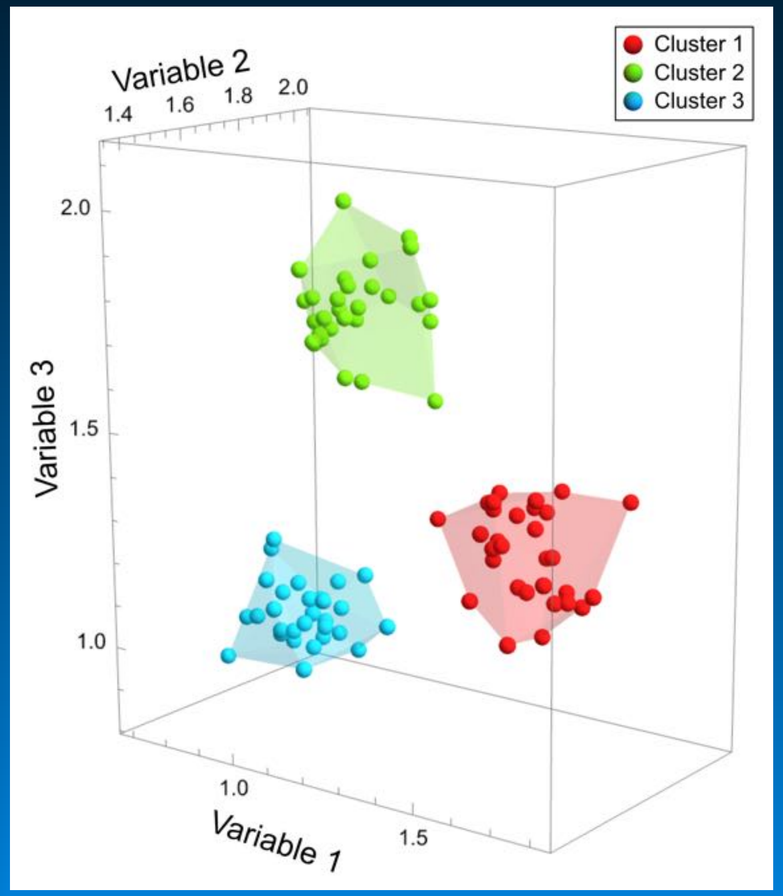


Optimal cluster solution (K = 3) for the example dataset.

Distribution of Points

How sensitive is this approach?

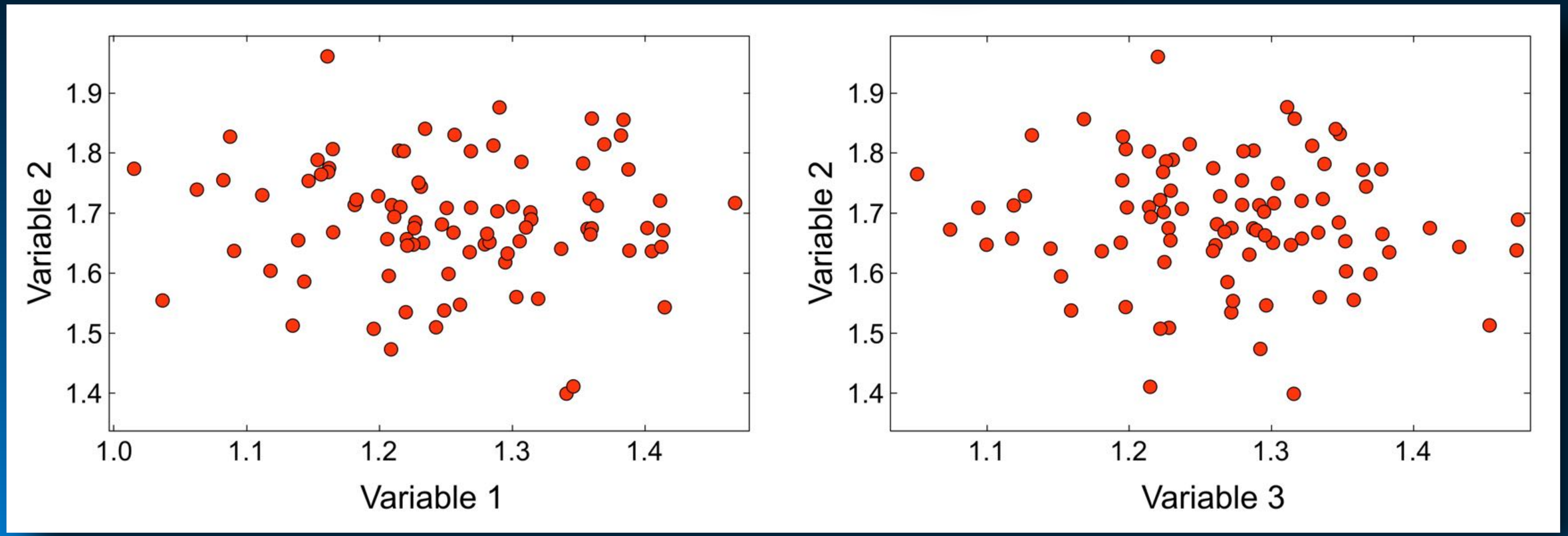




Optimal cluster solution (K = 3) for the example dataset.

Distribution of Points

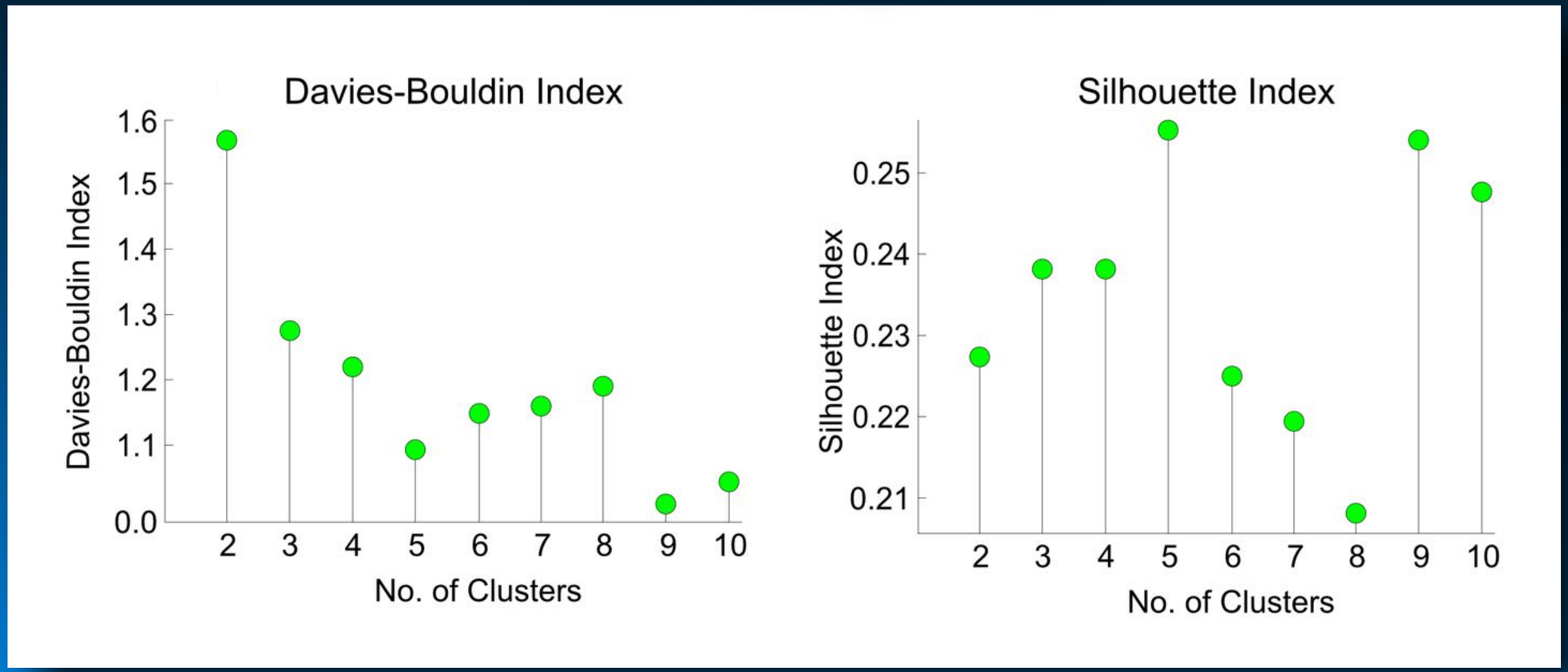
How sensitive is this approach?



Random Data

Distribution of Points

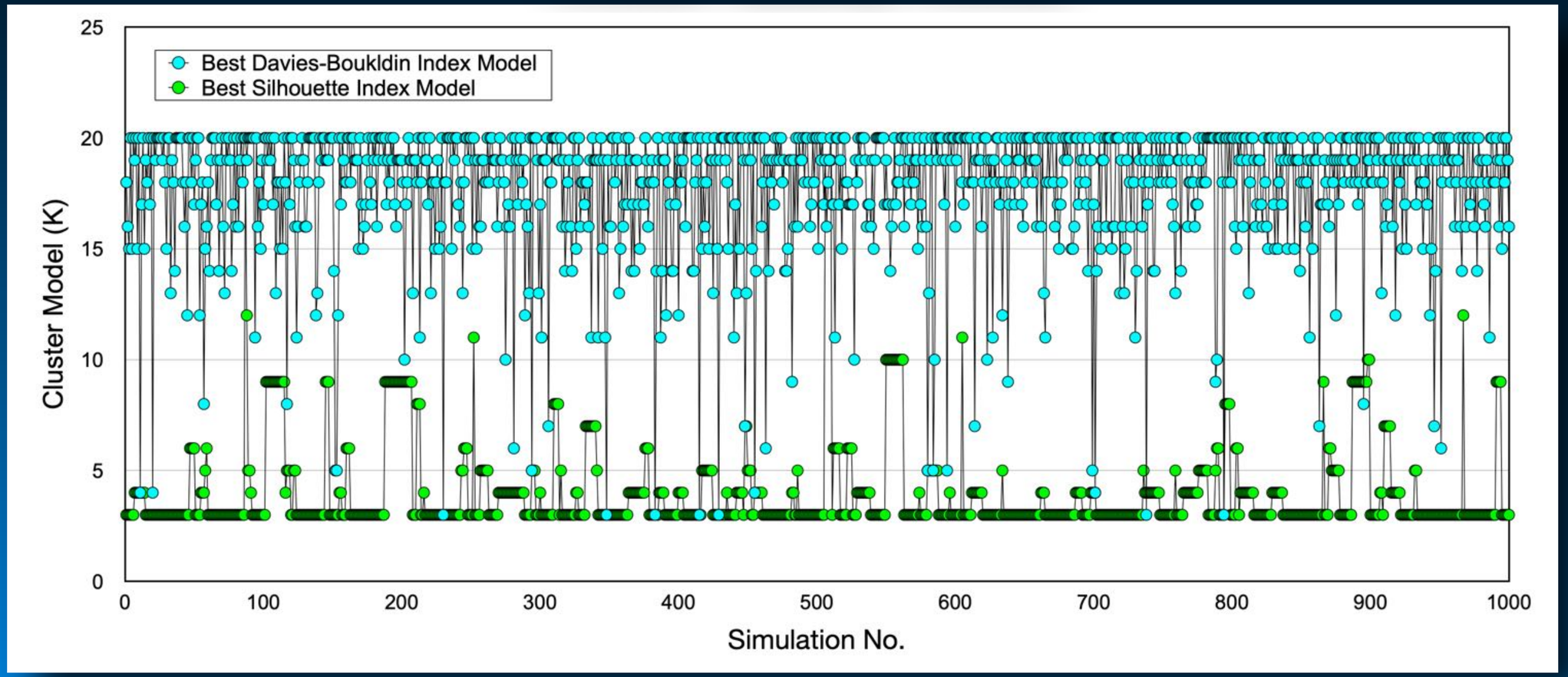
Cluster-Quality Indices



Note disagreement between the optimality states suggested by the Davies-Bouldin and Silhouette indices.

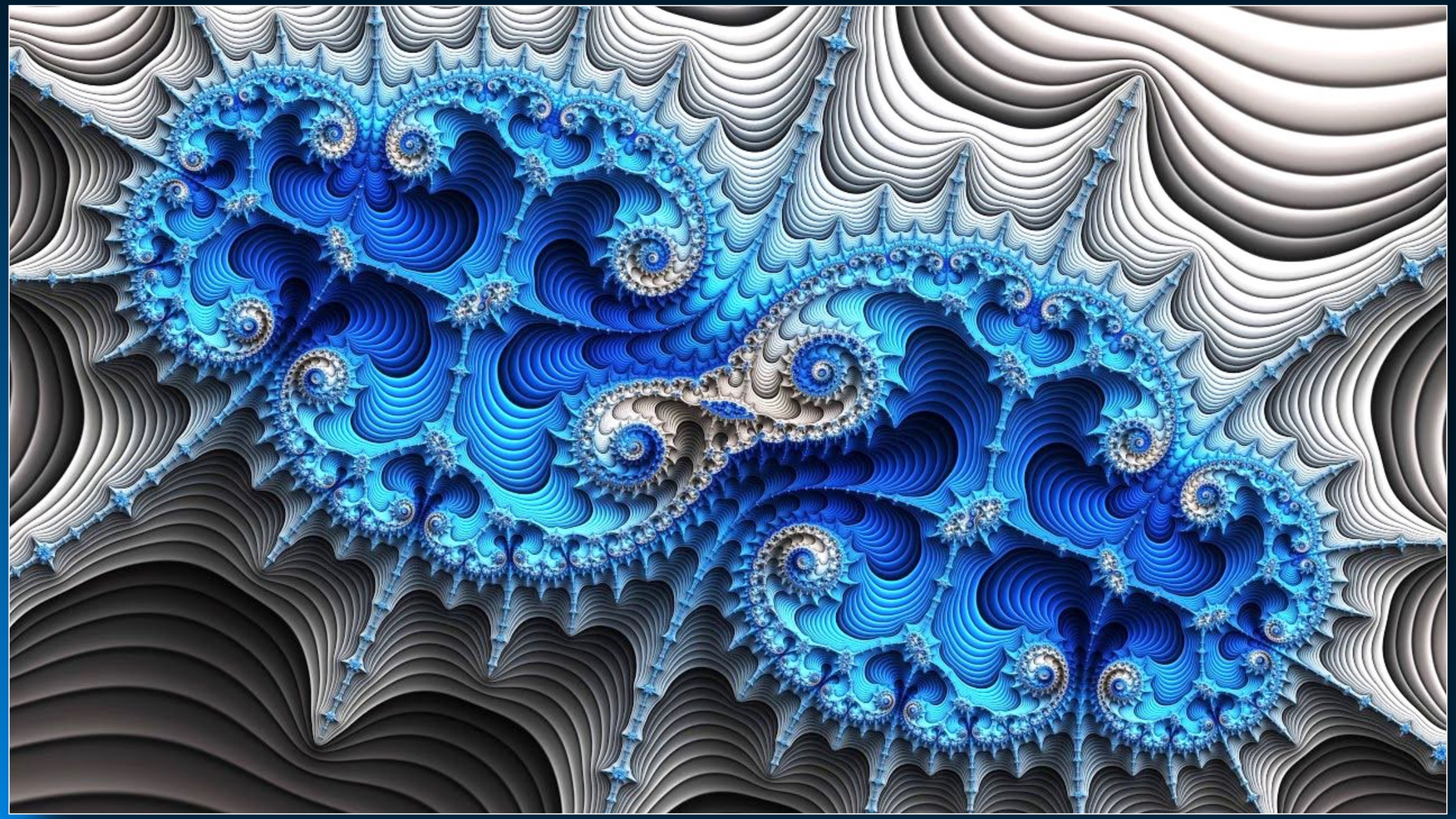
Distribution of Points

Cluster-Quality Indices



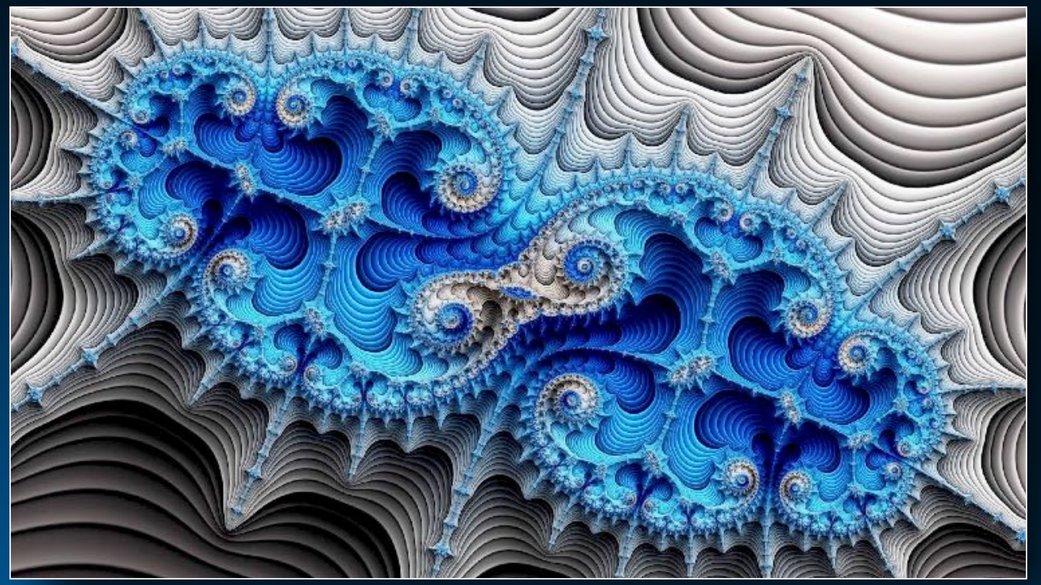
Results for 1,000 analyses of random (non-clustered) datasets. This approach yielded false positives in only 15 (0.015%) cases.

Fractal Analysis



Fractals

- Fractals are characteristic of some of the most geometrically complex objects known to science. And, since earth scientists often need to deal with, describe, and analyze geometrically complex objects – from molecules to entire planets – and understanding of fractals, and, methods to obtain fractal dimensions need to be part of every quantitative earth scientist's toolkit.

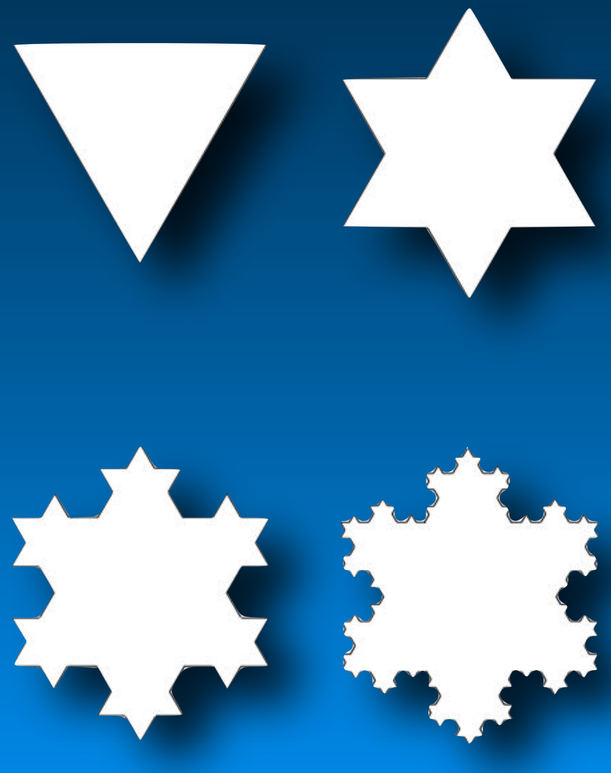


- The colloquial or conceptual definition of a fractal is any geometric line, surface or high-dimensional figure that exhibits a significant amount of "self similarity". This term is used to signify a geometric pattern that recurs at different scales across the form of two or higher dimensional objects.

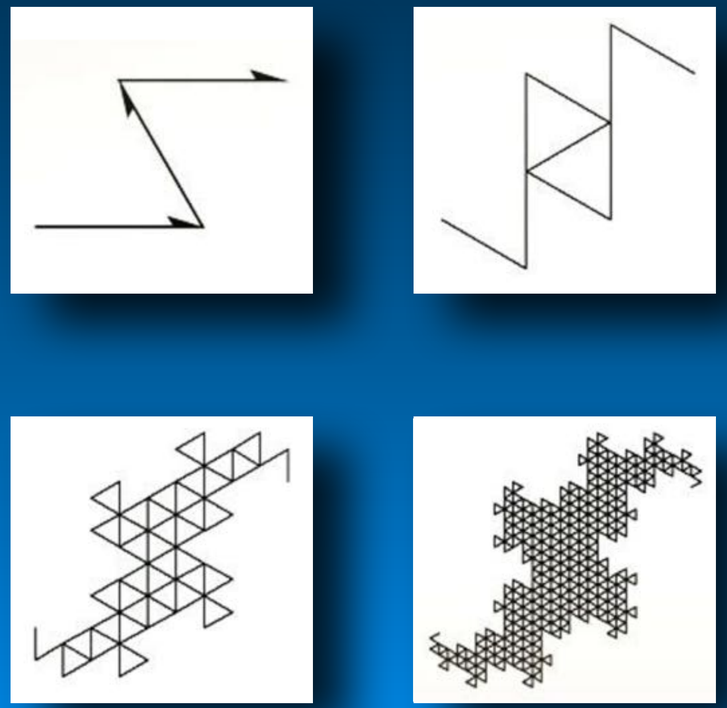
Fractals

Self-Similar Mathematical Objects

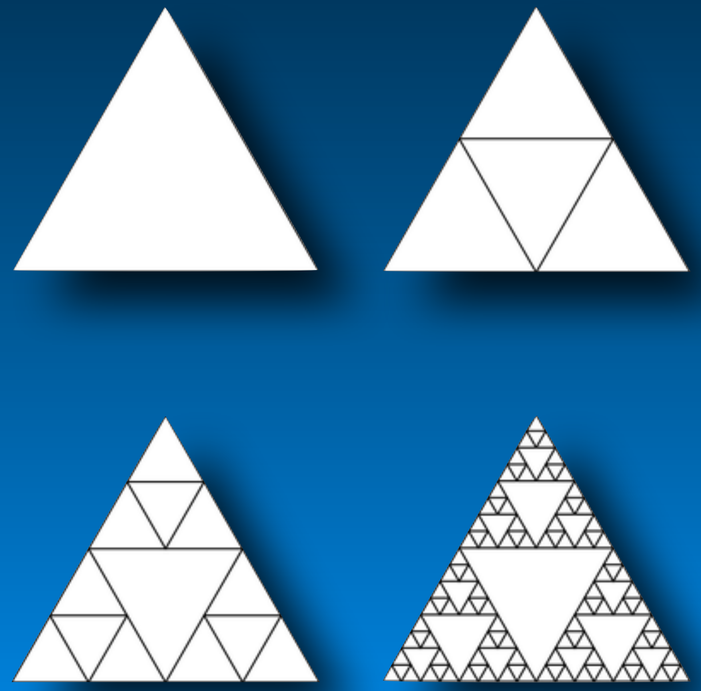
Koch Snowflake



Dragon Curve



Sierpinski's Triangle



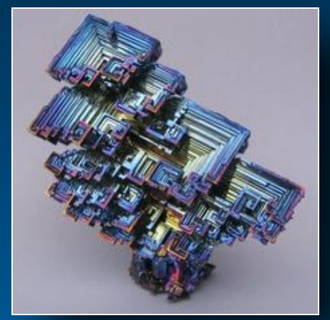
Fractals

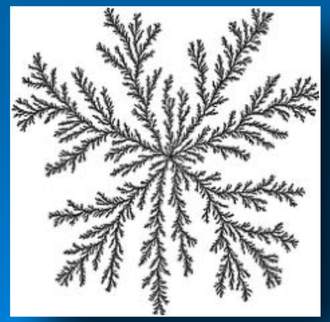
Self-Similar Earth Science Objects

Living Things




Minerals



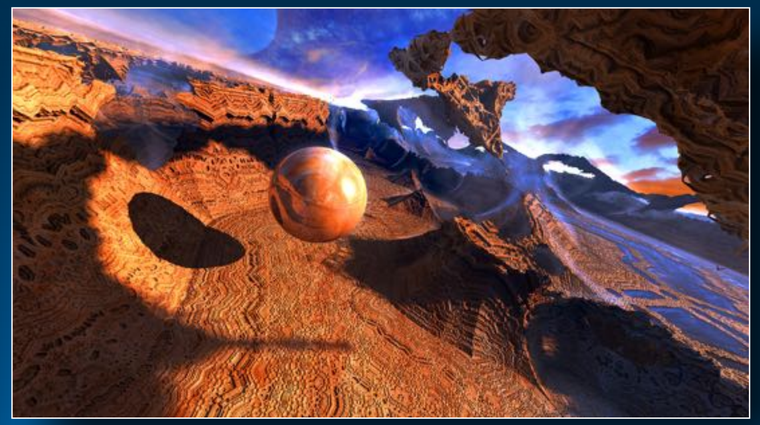



Structures




Fractals

Mathematically, fractals are even more complex. Indeed, theoretical mathematicians are still exploring the concept. At the same time, applied mathematicians and artists are using fractals to make ever more complex and “natural-looking” objects.




Fractal Analysis

Analytically. fractals are defined as any line or surface for which the Hausdorff-Besicovitch dimension exceeds its topological dimension.

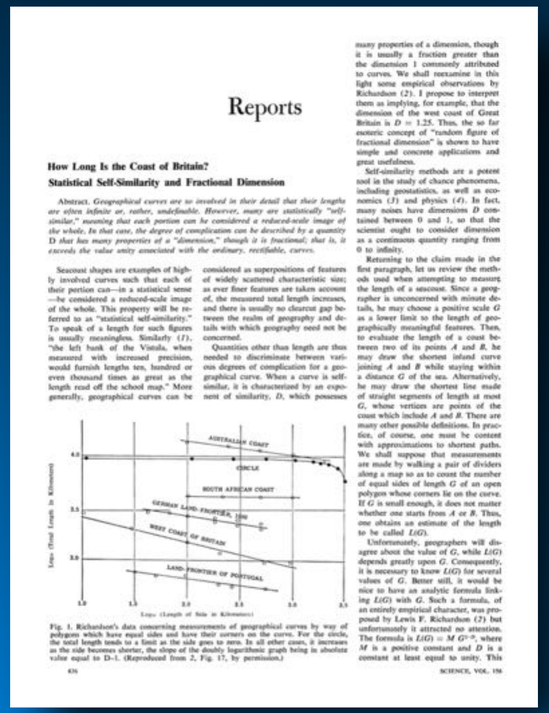
Lines are considered to have a single dimension because any point on the line can be located by a single number (distance from any other point). This holds for straight and simple curved lines.

Surfaces are considered to have a two dimensions because any point on the surface can be located by a two numbers (x and y distance from any other point). This holds for planar and uniformly curving surfaces.

Fractals have an intermediate character because their geometry is characterized by regular (= self similar) irregularities at every scale. Hence, distances between points are a function of scale.

Fractal Analysis

The classic fractal problem is an earth-science question: How long is the coast of Britain?



Reports

How Long Is the Coast of Britain?
Statistical Self-Similarity and Fractional Dimension

Abstract. *Geographical curves are so involved in their detail that their lengths are often infinite or, rather, undefinable. However, many are statistically "self-similar," meaning that each portion can be considered a reduced-scale image of the whole. In that case, the degree of complication can be described by a quantity D that has many properties of a "dimension," though it is fractional; that is, it exceeds the value unity associated with the ordinary, rectifiable, curves.*

Seacoast shapes are examples of highly involved curves such that each of their portion can—in a statistical sense—be considered a reduced-scale image of the whole. This property will be referred to as "statistical self-similarity." To speak of a length for such figures is usually meaningless. Similarly (*1*), "the left bank of the Vistula, when measured with increased precision, would furnish lengths ten, hundred or even thousand times as great as the length read off the school map." More generally, geographical curves can be considered as superpositions of features of widely scattered characteristic size; as ever finer features are taken account of, the measured total length increases, and there is usually no clearcut gap between the realm of geography and details with which geography need not be concerned.

Quantities other than length are thus needed to discriminate between various degrees of complication for a geographical curve. When a curve is self-similar, it is characterized by an exponent of similarity, *D*, which possesses many properties of a dimension, though it is usually a fraction greater than the dimension 1 commonly attributed to curves. We shall reexamine in this light some empirical observations by Richardson (*2*). I propose to interpret them as implying, for example, that the dimension of the west coast of Great Britain is $D \approx 1.25$. Thus, the so far esoteric concept of "random figure of fractional dimension" is shown to have simple and concrete applications and great usefulness.

Self-similarity methods are a potent tool in the study of chance phenomena, including geostatistics, as well as economics (*3*) and physics (*4*). In fact, many noises have dimensions *D* contained between 0 and 1, so that the scientist ought to consider dimension as a continuous quantity ranging from 0 to infinity.

Returning to the claim made in the first paragraph, let us review the methods used when attempting to measure the length of a seacoast. Since a geographer is unconcerned with minute details, he may choose a positive scale *G* as a lower limit to the length of geographically meaningful features. Then, to evaluate the length of a coast between two of its points *A* and *B*, he may draw the shortest inland curve joining *A* and *B* while staying within a distance *G* of the sea. Alternatively, he may draw the shortest line made of straight segments of length at most *G*, whose vertices are points of the coast which include *A* and *B*. There are many other possible definitions. In practice, of course, one must be content with approximations to shortest paths. We shall suppose that measurements are made by walking a pair of dividers along a map so as to count the number of equal sides of length *G* of an open polygon whose corners lie on the curve. If *G* is small enough, it does not matter whether one starts from *A* or *B*. Thus, one obtains an estimate of the length to be called *L*(*G*).

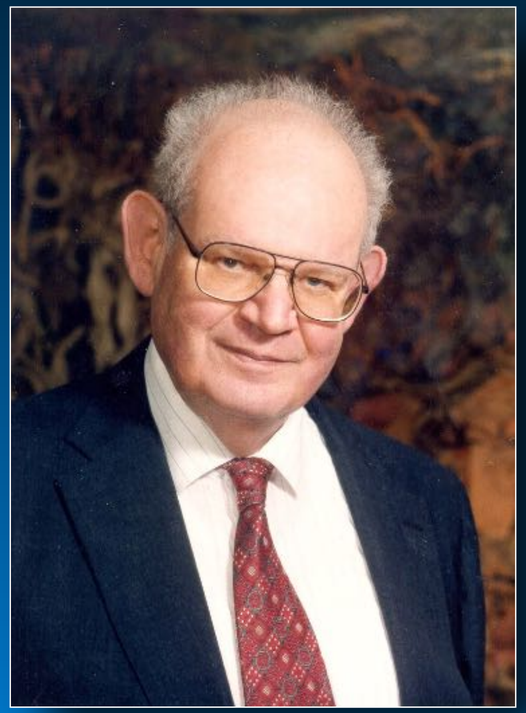
Unfortunately, geographers will disagree about the value of *G*, while *L*(*G*) depends greatly upon *G*. Consequently, it is necessary to know *L*(*G*) for several values of *G*. Better still, it would be nice to have an analytic formula linking *L*(*G*) with *G*. Such a formula, of an entirely empirical character, was proposed by Lewis F. Richardson (*2*) but unfortunately it attracted no attention. The formula is $L(G) \approx M\,G^{1-D}$, where *M* is a positive constant and *D* is a constant at least equal to unity. This

Fig. 1. Richardson's data concerning measurements of geographical curves by way of polygons which have equal sides and have their corners on the curve. For the circle, the total length tends to a limit as the side goes to zero. In all other cases, it increases as the side becomes shorter, the slope of the doubly logarithmic graph being in absolute value equal to D-1. (Reproduced from 2, Fig. 17, by permission.)

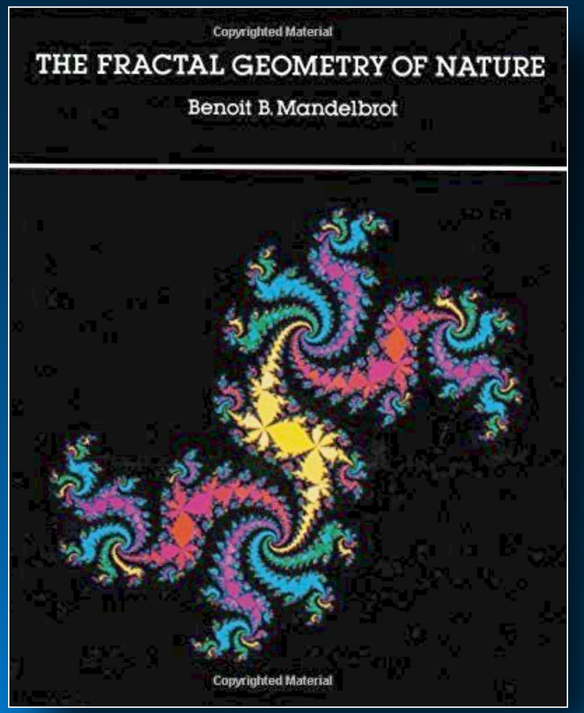
636

SCIENCE, VOL. 156

Mandelbrot, B.B., 1967, How long is the coast of Britain? Statistical self-similarity and fractional dimension: Science, v. 155, p. 636–638.



Benoit Mandelbrot
(1924 – 2010)

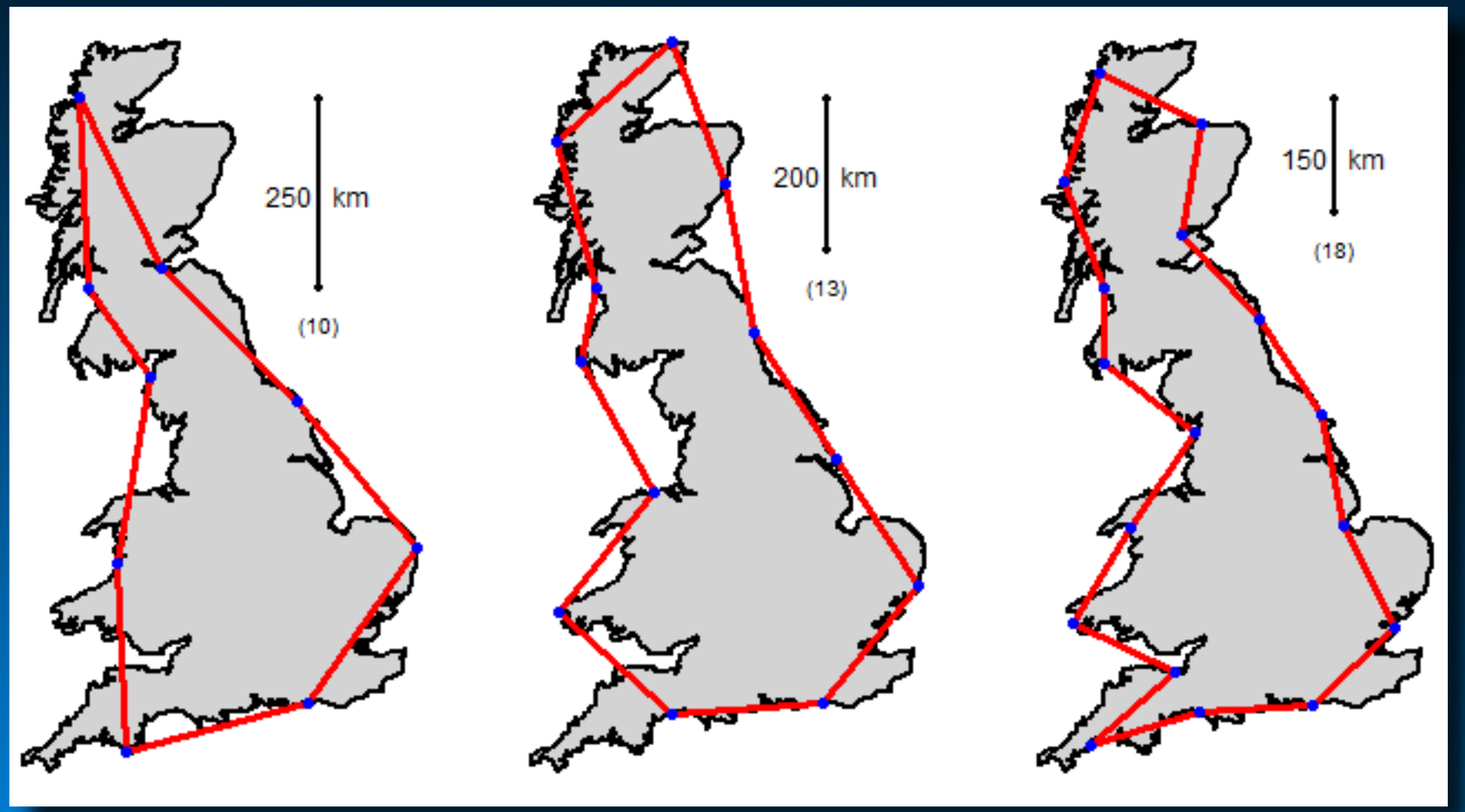


Mandelbrot, B.B., 1982, The fractal geometry of nature: San Francisco, California, W. H. Freeman & Company Limited, 460 p.

Fractal Analysis

How long is the coast of Britain?

What's the length of the coastline?



10 segments
250 km/segment
2,500 km

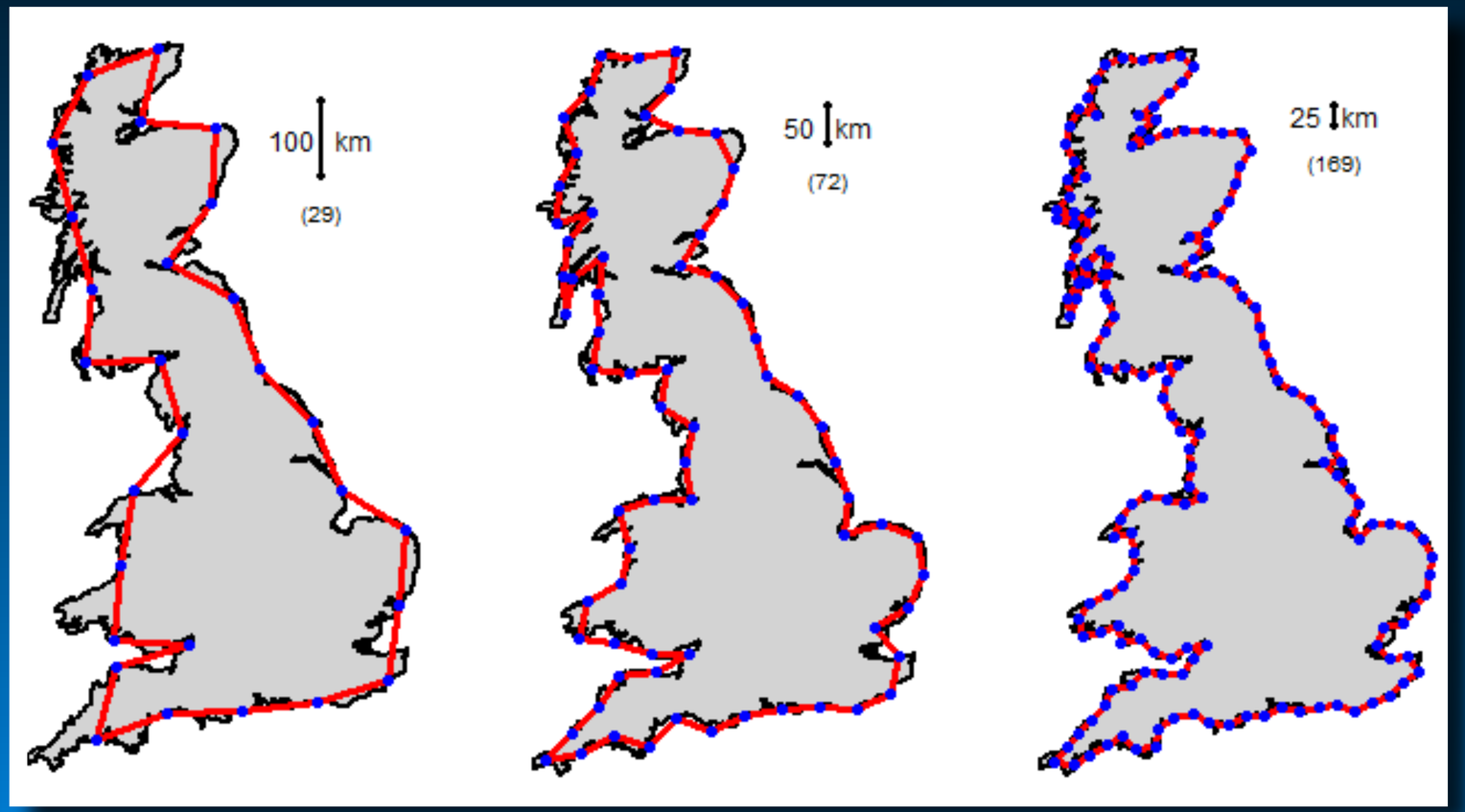
13 segments
200 km/segment
2,600 km

18 segments
150 km/segment
2,700 km

Fractal Analysis

How long is the coast of Britain?

What's the length of the coastline?



29 segments
100 km/segment
2,900 km

50 segments
72 km/segment
3,600 km

169 segments
25 km/segment
4,225 km

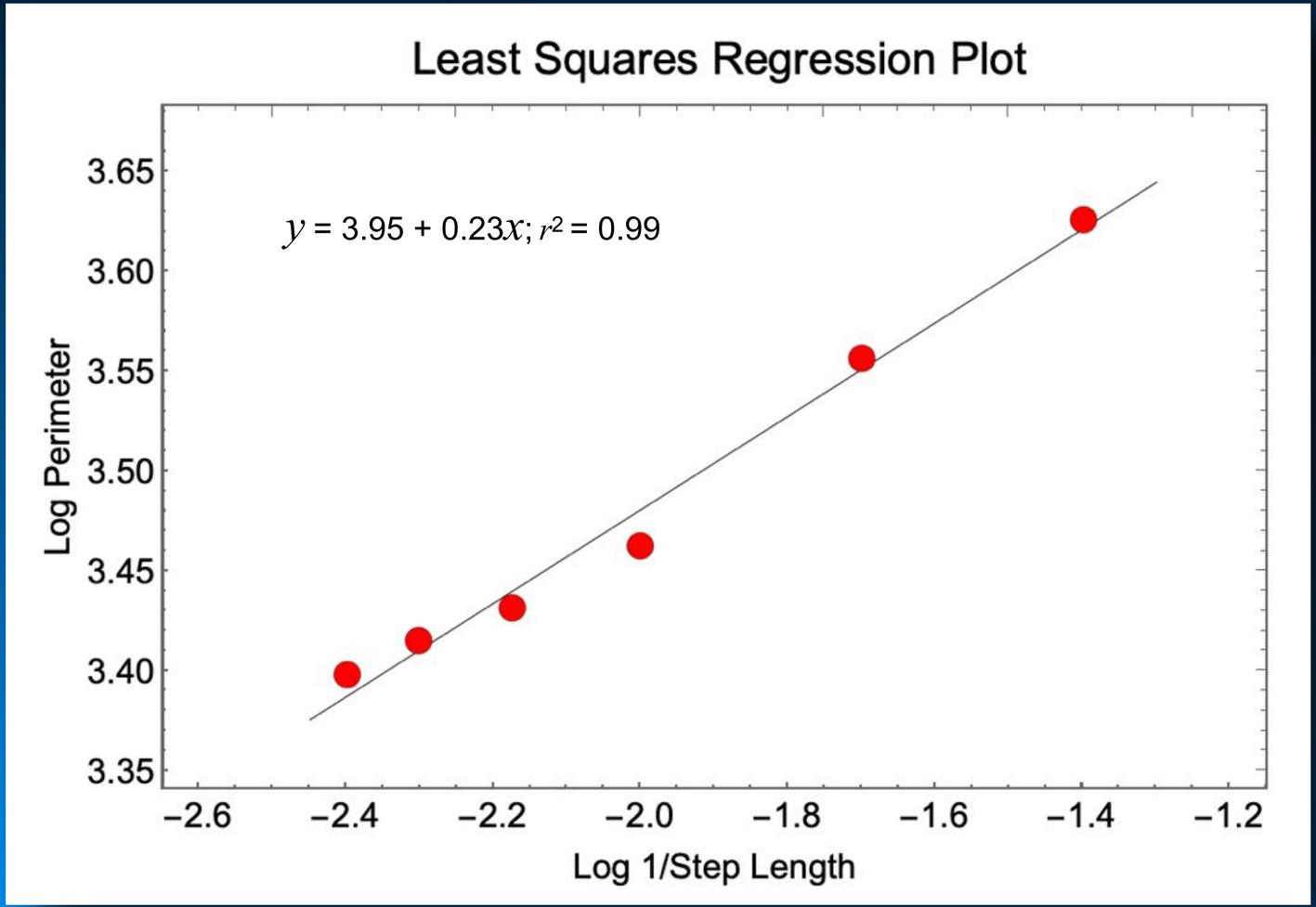
Fractal Analysis

How long is the coast of Britain?

Observation	No of Steps	Step Length	1/Step Length	Perimeter
1	10	250	0.004	2,500
2	13	200	0.005	2,600
3	18	150	0.007	2,700
4	29	100	0.010	2,900
5	72	50	0.020	3,600
6	169	25	0.040	4,225

Fractal Analysis

How long is the coast of Britain?



Fractal Analysis

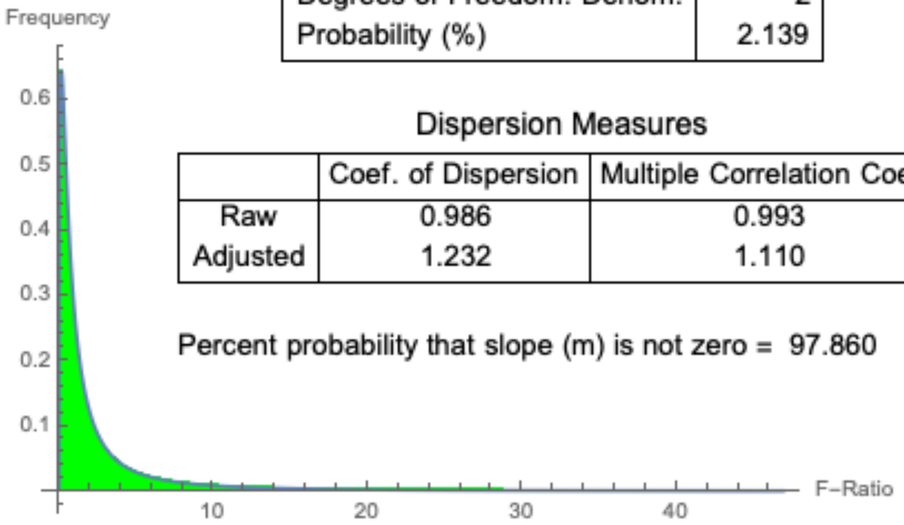
Coast of Britain Regression Statistics

Regression ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Regression	0.040	3	0.013	45.910
Deviation	0.001	2	0.000	
Total	0.041	5		

Probability Table

Observed F−value	45.910
Degrees of Freedom: Num.	3
Degrees of Freedom: Denom.	2
Probability (%)	2.139



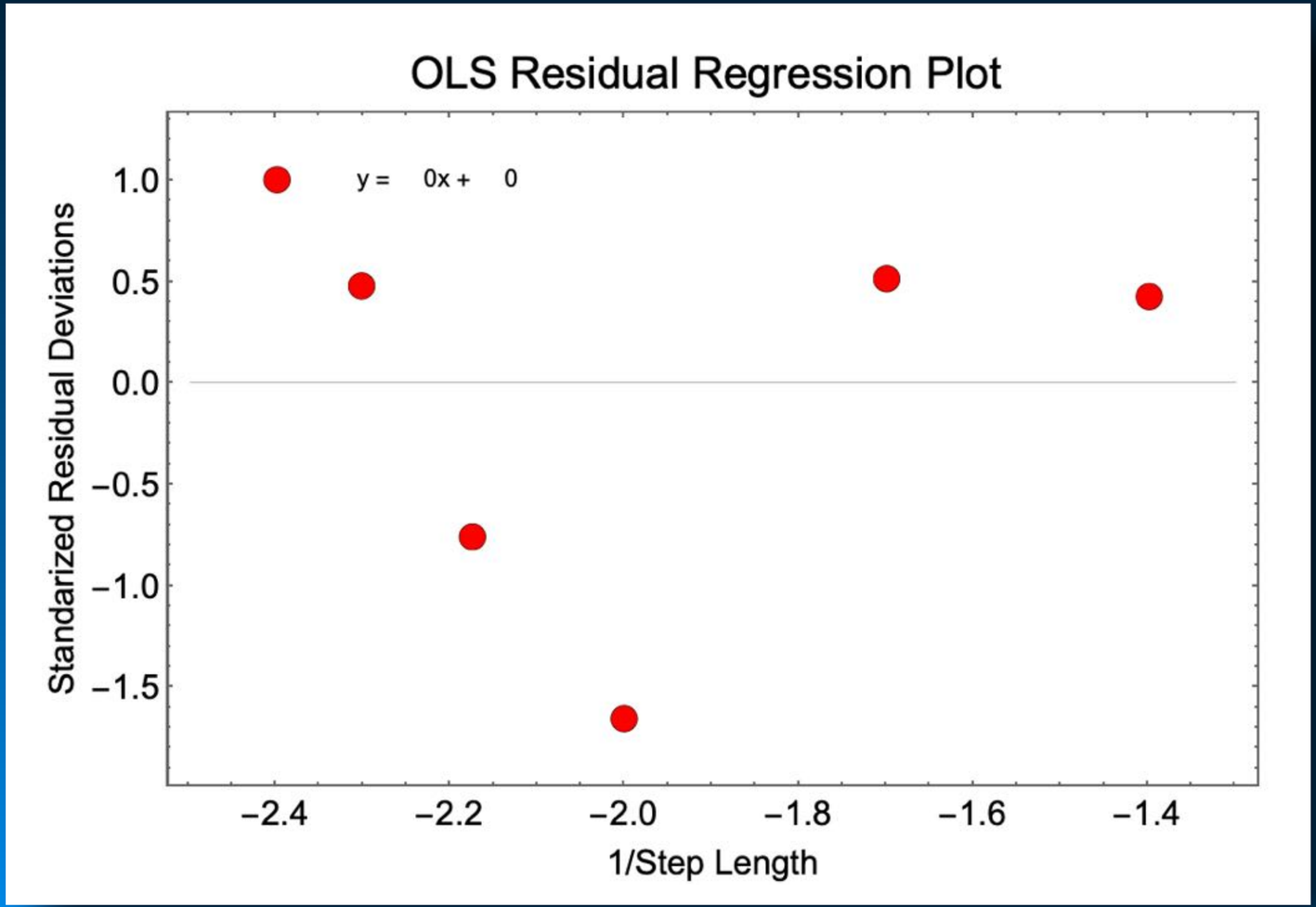
Dispersion Measures

	Coef. of Dispersion	Multiple Correlation Coef.
Raw	0.986	0.993
Adjusted	1.232	1.110

Percent probability that slope (m) is not zero = 97.860

Fractal Analysis

Coast of Britain Regression Statistics



Fractal Analysis

How to Convert the Regression Result into a Fractal Dimension

$$y = 3.95 + 0.23x$$

$$log\ y\ = 3.95 + 0.23\ \cdot log\ x$$

$$log\ nr = 3.95 + 0.23 \cdot log\ (\frac{1}{r})$$

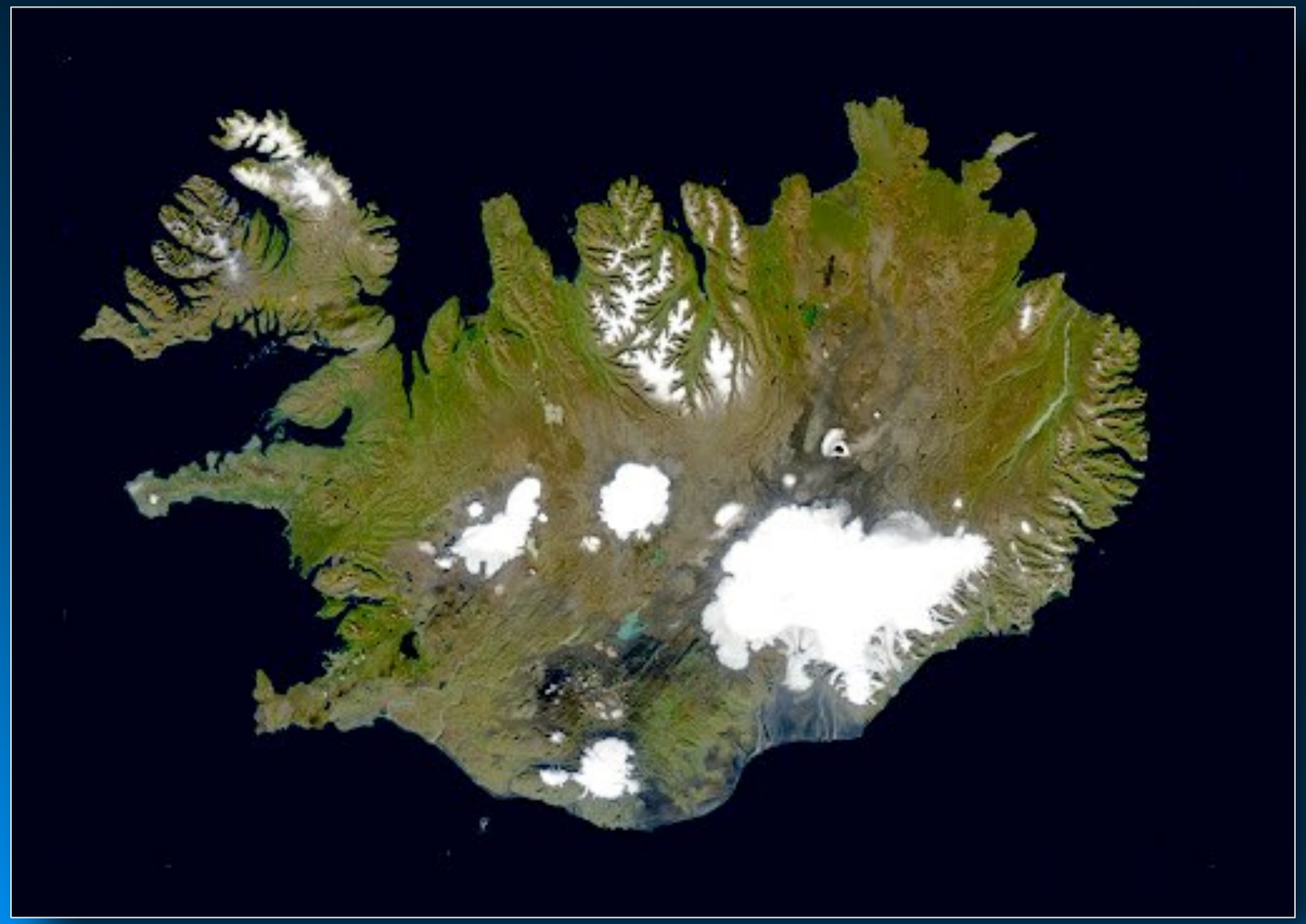
$$log\ Perimeter = 3.95 + 0.23\ log\ (\frac{1}{Step\ Length})$$

$$Perimeter = 3.95 \cdot (\frac{1}{Step\ Length})^{0.23}$$

$$D_f = 1 + \beta = 1.23$$

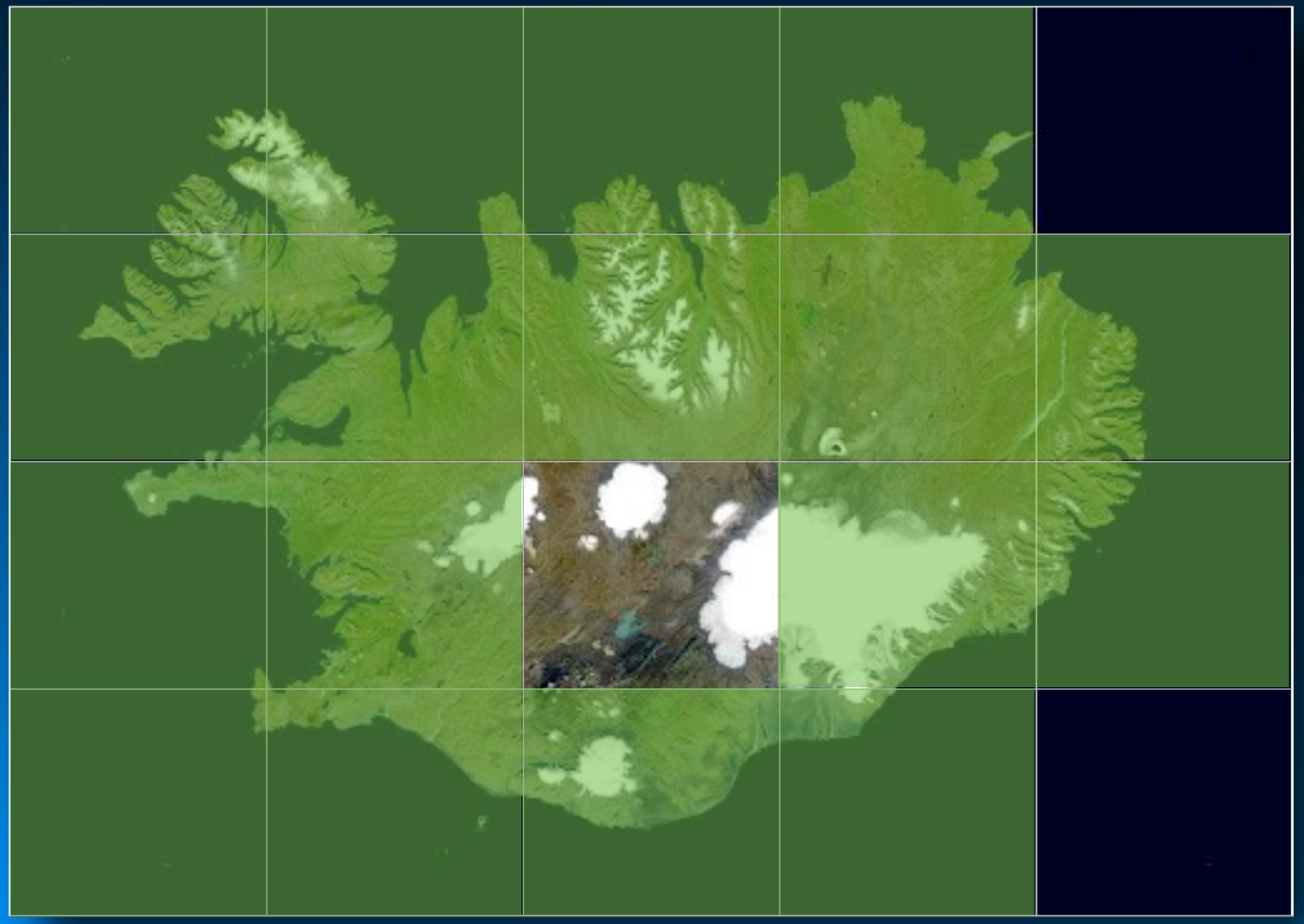
Fractal Analysis

An Alternative Approach to Calculating D_f



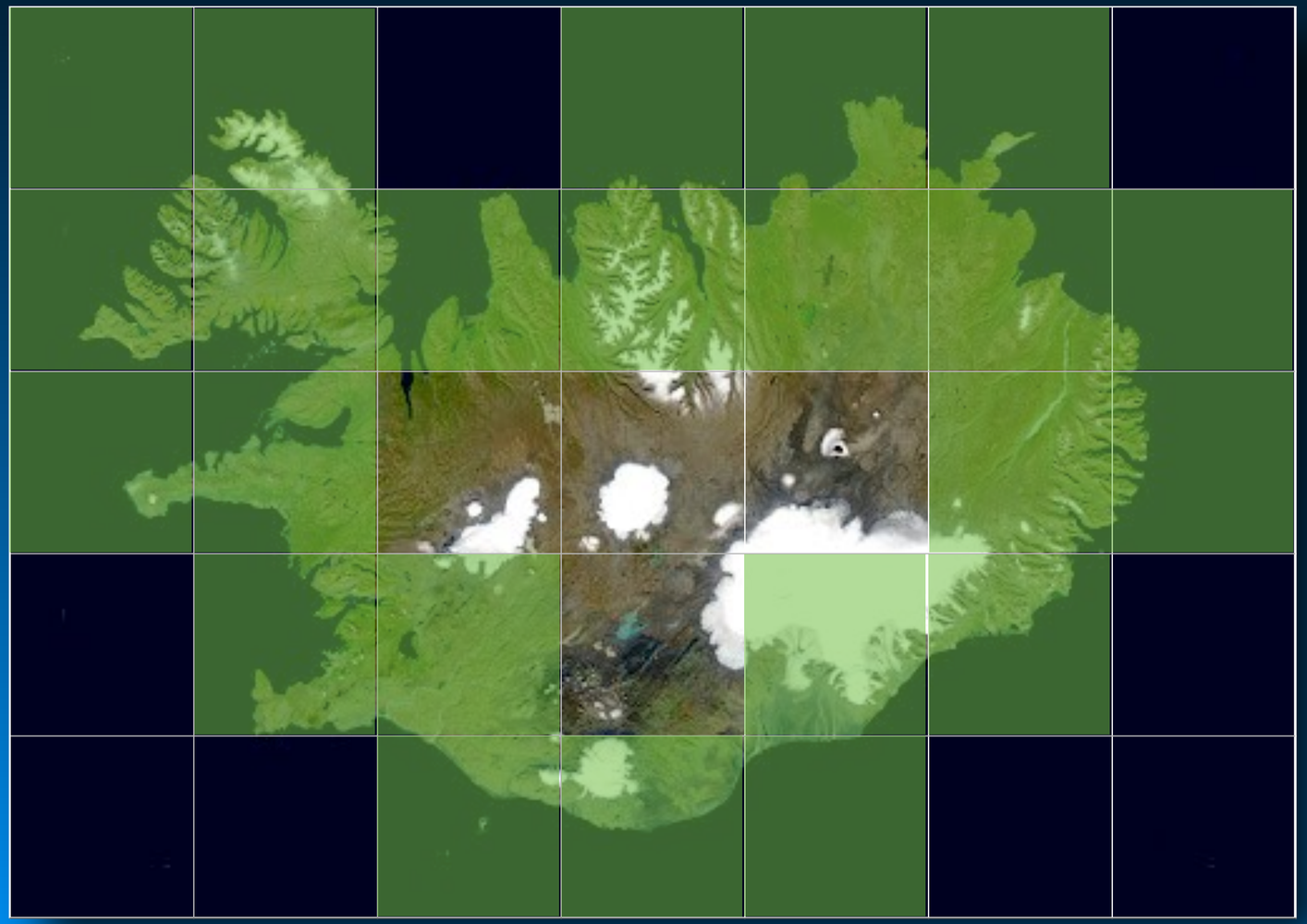
Fractal Analysis

An Alternative Approach to Calculating D_f



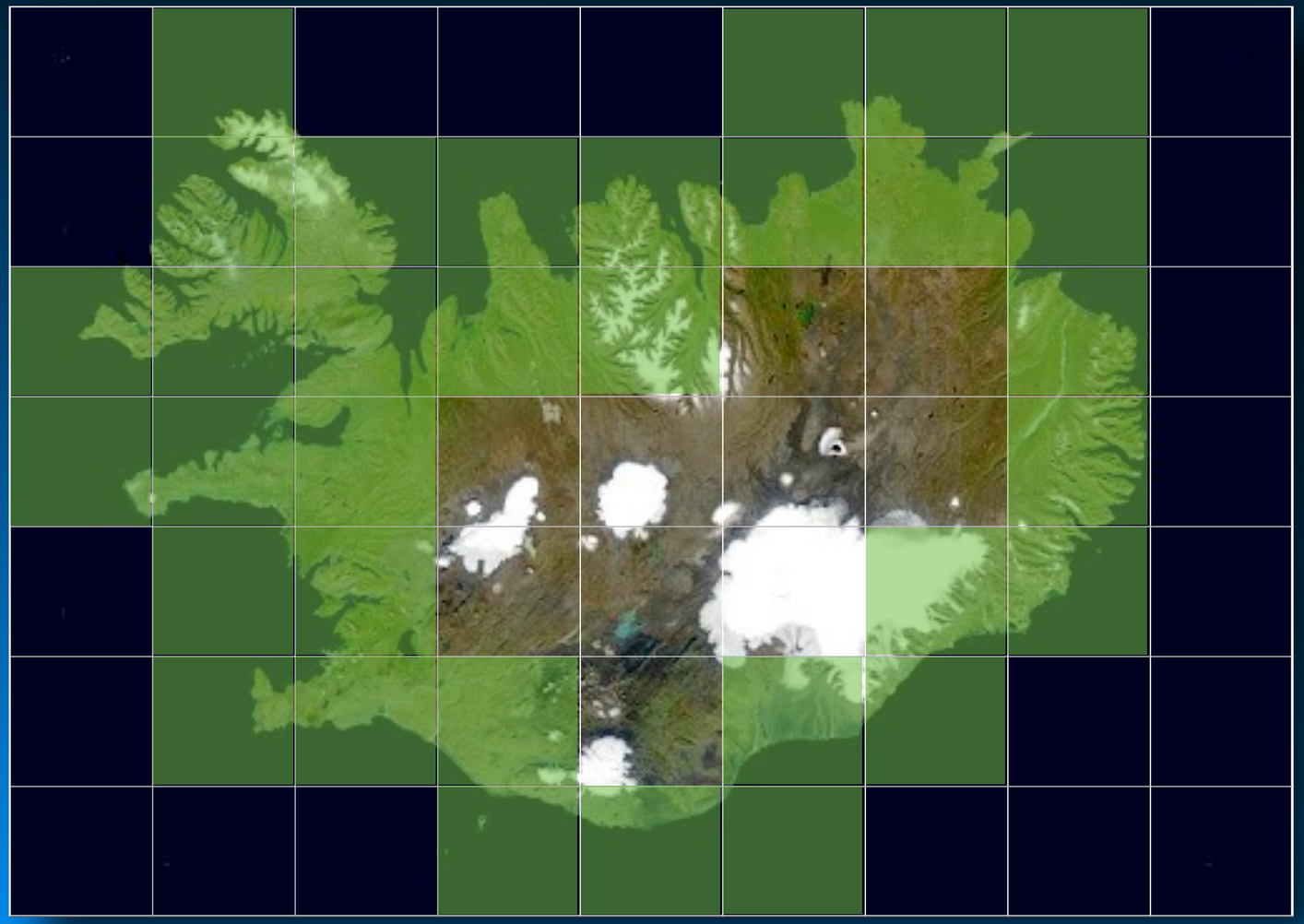
Fractal Analysis

An Alternative Approach to Calculating D_f



Fractal Analysis

An Alternative Approach to Calculating D_f



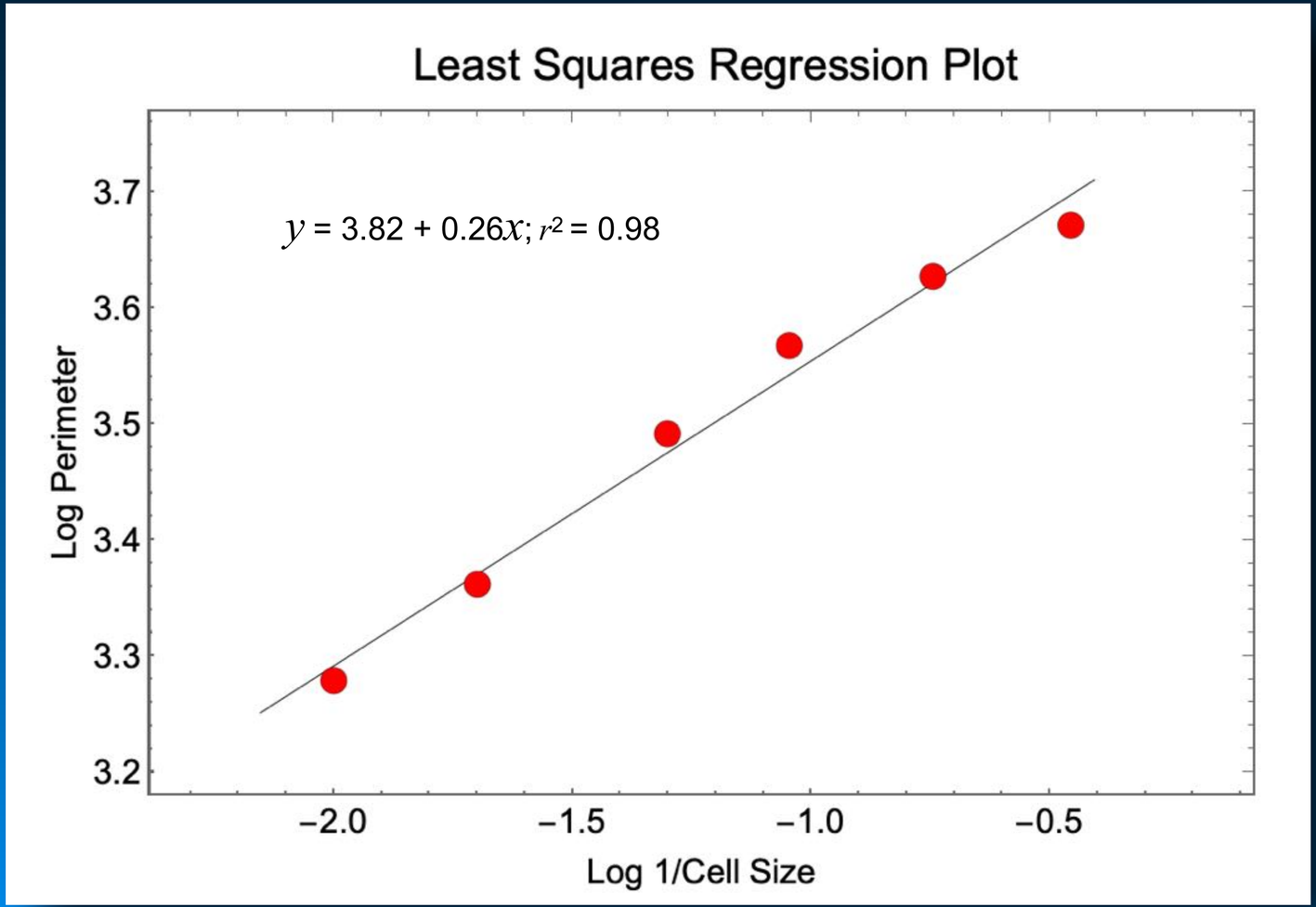
Fractal Analysis

How long is the coast of Britain?

Observation	No of Cells	Cell Size	1/Cell Size	Perimeter
1	19	100	0.010	1,900
2	46	50	0.020	2,300
3	155	20	0.050	3,100
4	321	11.5	0.087	3,692
5	743	5.7	0.175	4,235
6	1,645	2.85	0.351	4,688

Fractal Analysis

An Alternative Approach to Calculating D_f



Fractal Analysis

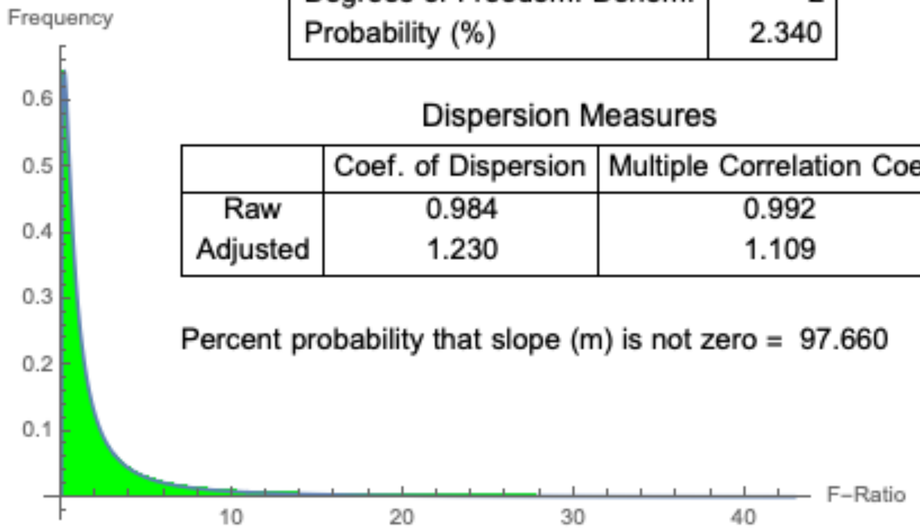
Coast of Iceland Regression Statistics

Regression ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Regression	0.116	3	0.039	41.896
Deviation	0.002	2	0.001	
Total	0.118	5		

Probability Table

Observed F−value	41.896
Degrees of Freedom: Num.	3
Degrees of Freedom: Denom.	2
Probability (%)	2.340



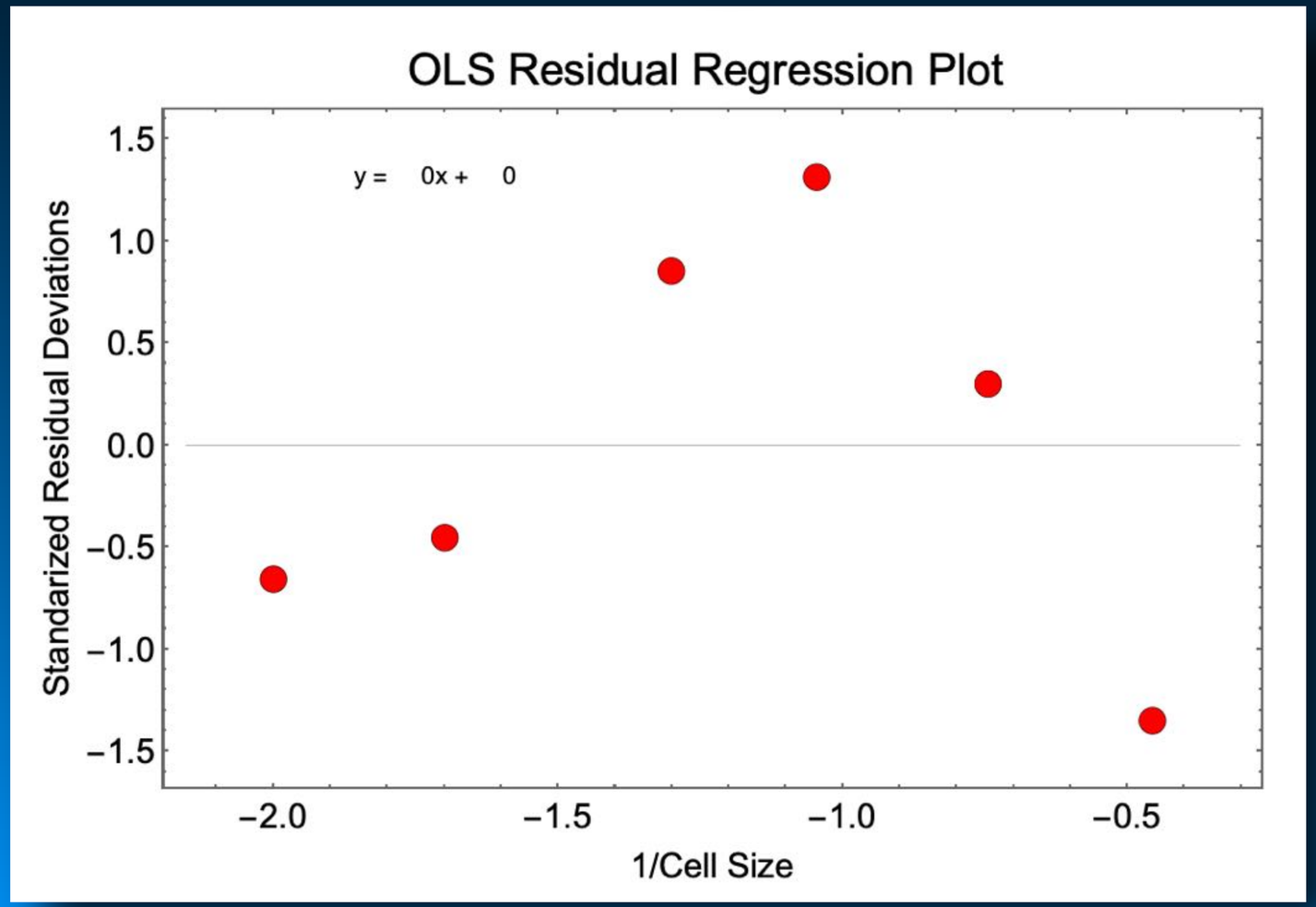
Dispersion Measures

	Coef. of Dispersion	Multiple Correlation Coef.
Raw	0.984	0.992
Adjusted	1.230	1.109

Percent probability that slope (m) is not zero = 97.660

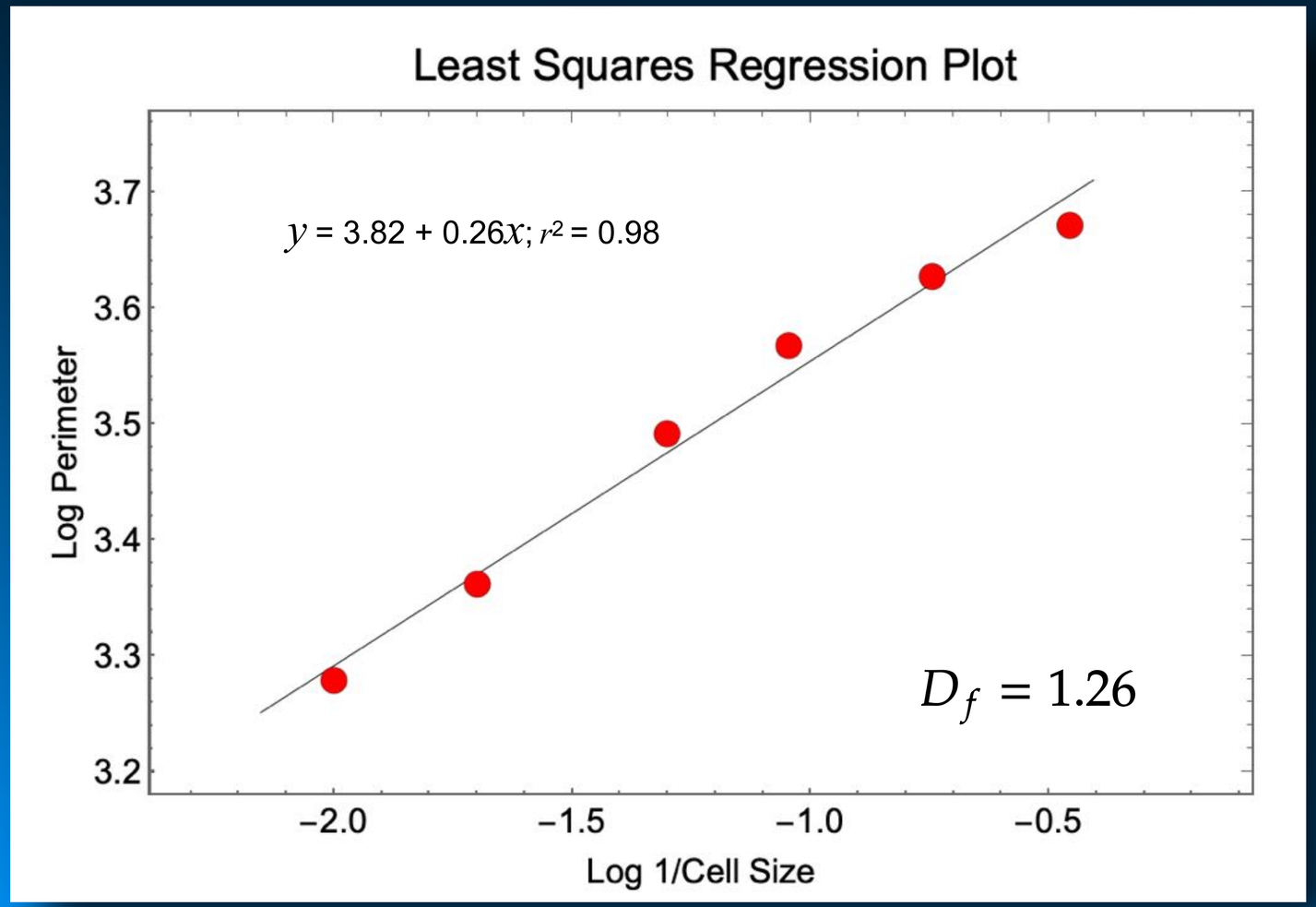
Fractal Analysis

Coast of Iceland Regression Statistics



Fractal Analysis

An Alternative Approach to Calculating D_f

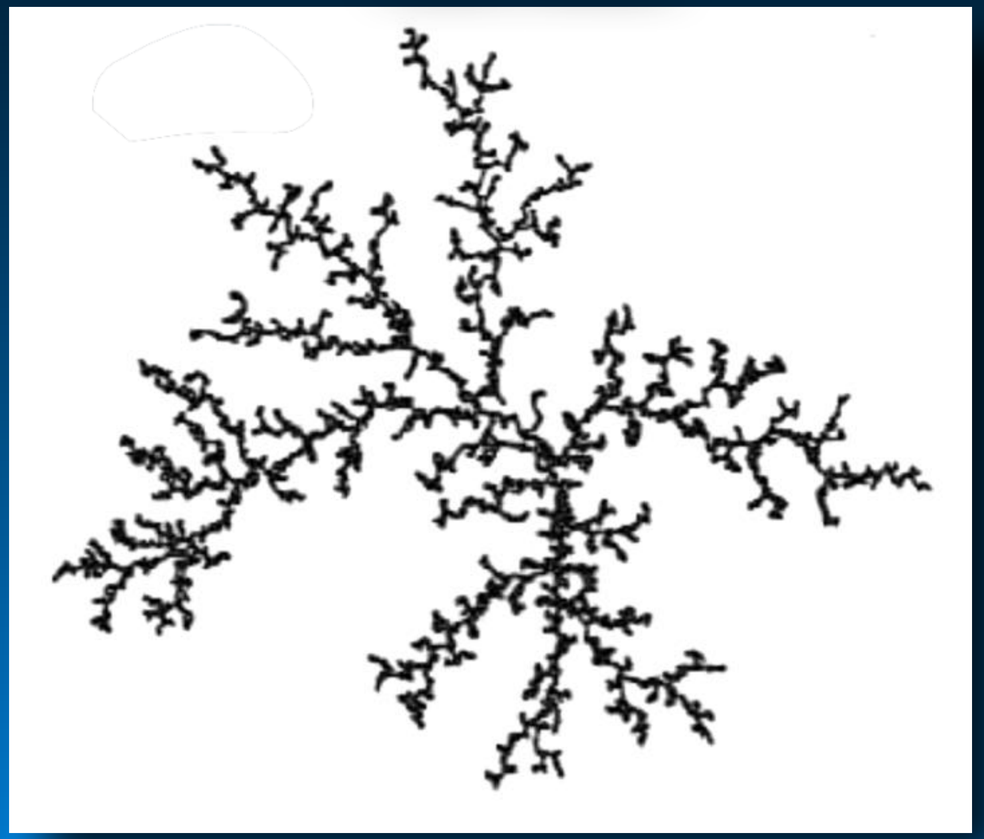


Fractal Analysis

Fracture Pattern in Sandstone

$D_f = 1.52$

Pattern of Porosity in Packed Glass Beads

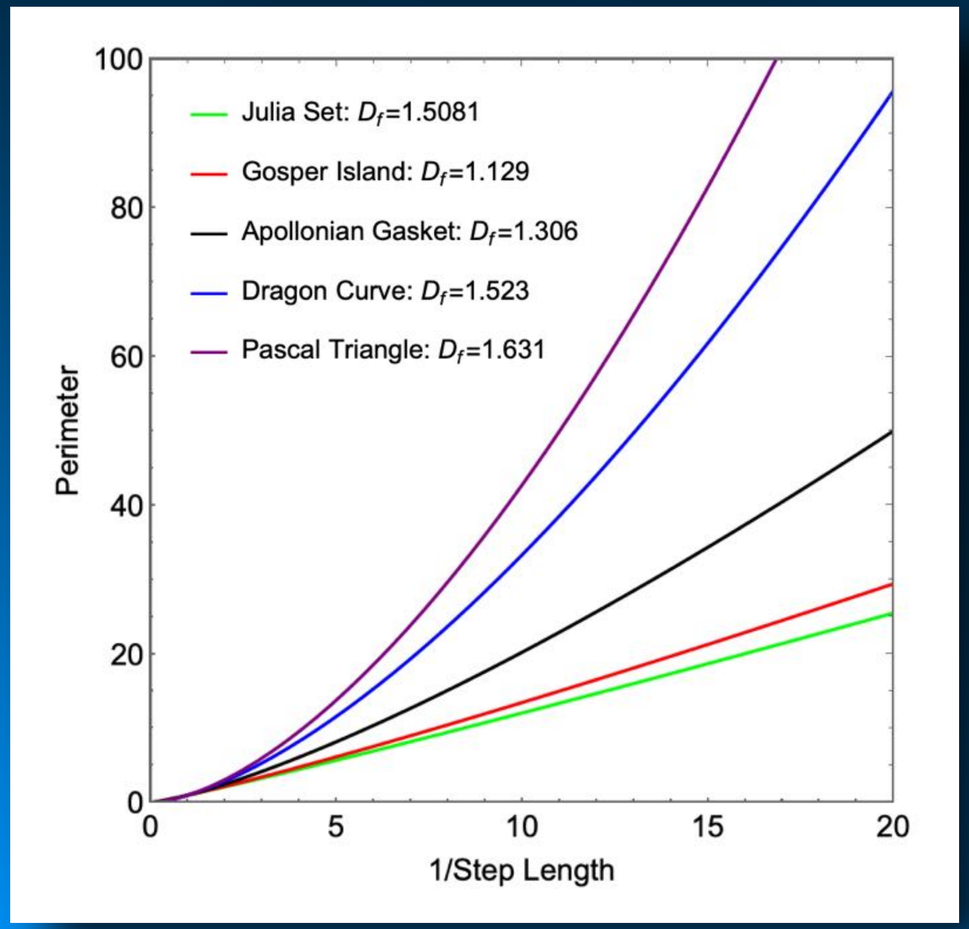


$D_f = 1.64$

Fractal Analysis

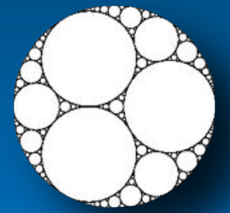
Practical & Theoretical Fractals

Math Function Fractals

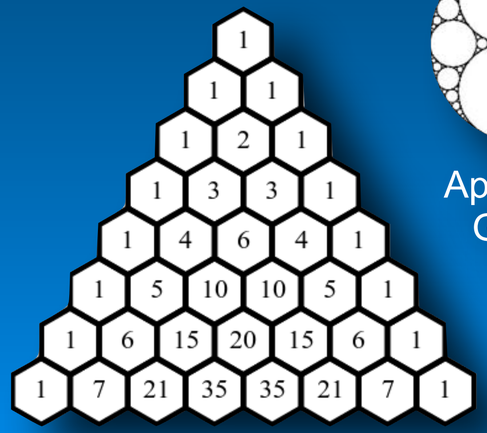


Julia Set

Gosper Island



Apollonian Gasket



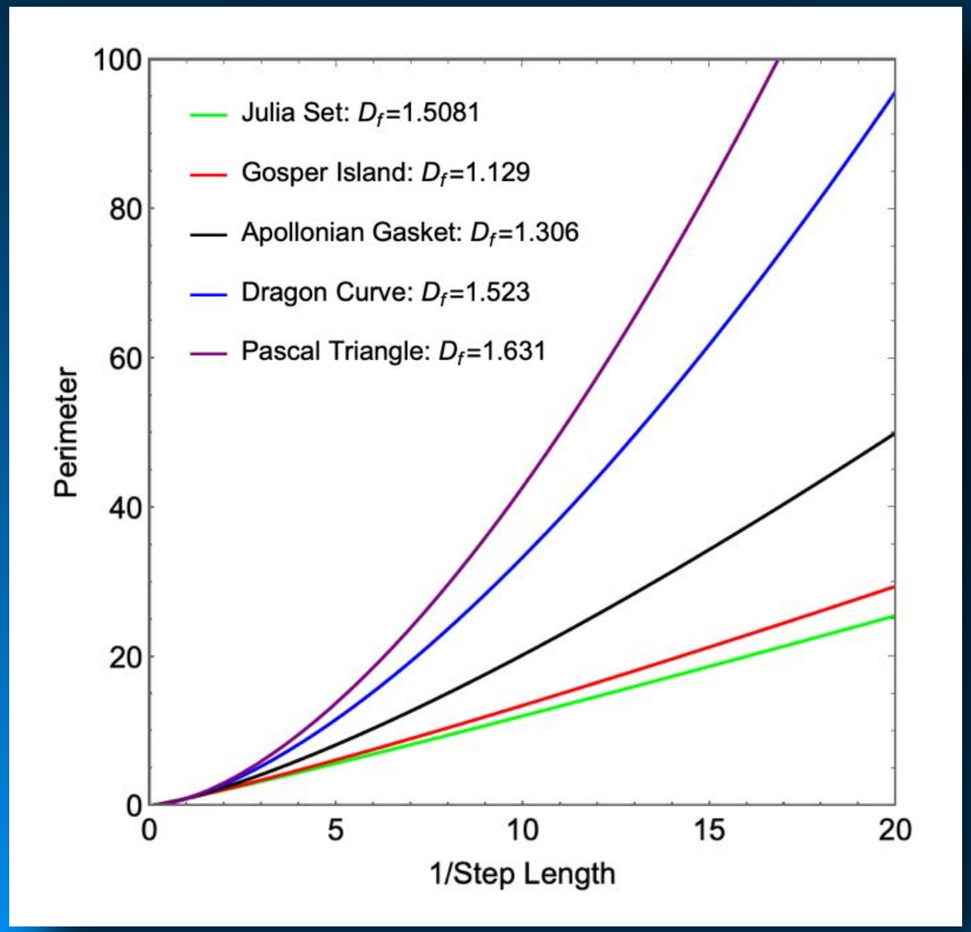
Pascal Triangle

Dragon Curve

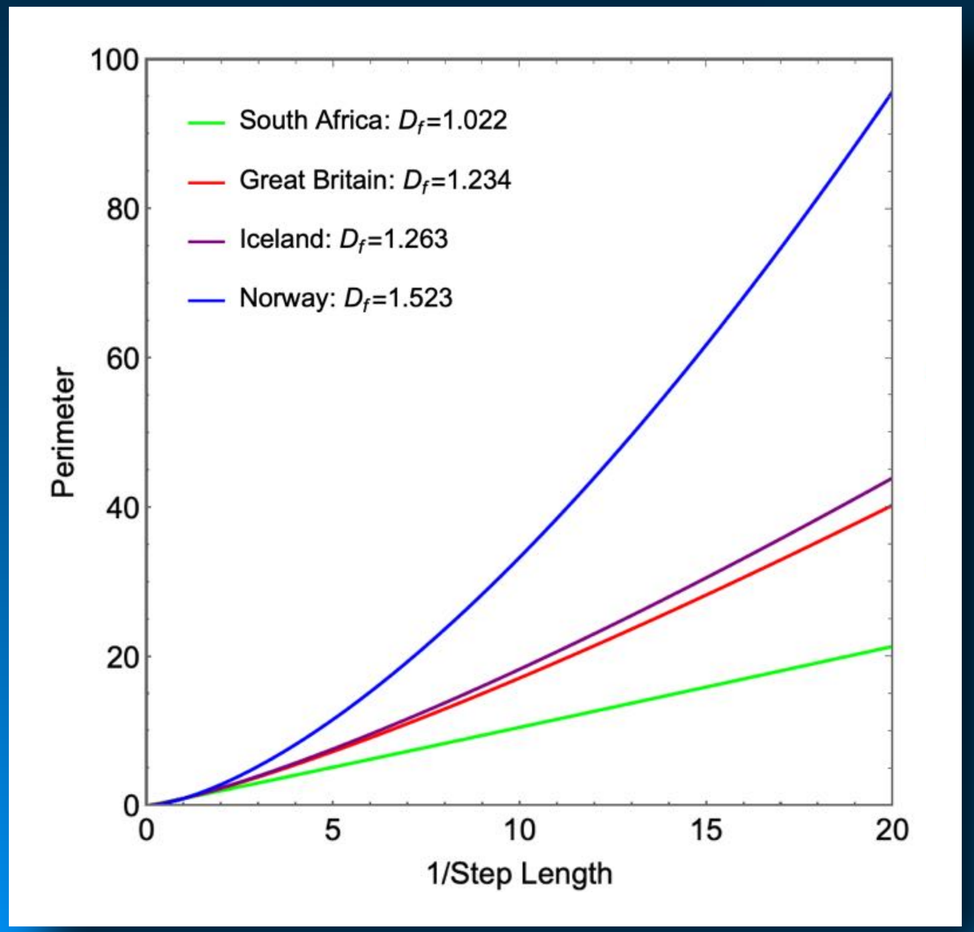
Fractal Analysis

Practical & Theoretical Fractals

Math Function Fractals



Coastline Fractals



Fractals

Any line or surface for which the Hausdorff-Besicovitch dimension exceeds its topological dimension.

Advantages

- Simple model.
- Quick to calculate.
- Allows complex patterns to be represented quantitatively, compared and tested against hypothesis-driven expectations.

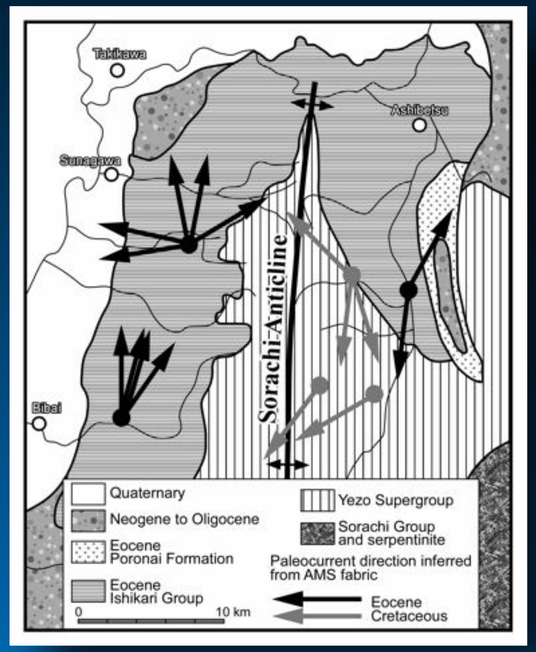
Disadvantages

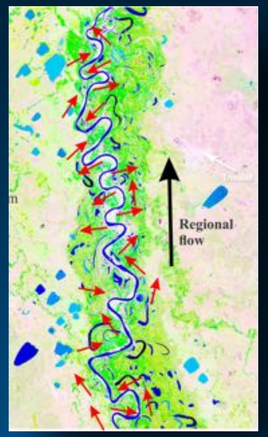
- Fractals are descriptions of one aspect of the systems they are used to describe, the do not represent an adequate description of the geometry as a whole.
- Comparisons between fractal dimensions characterizing different samples are difficult to interpret geometrically.

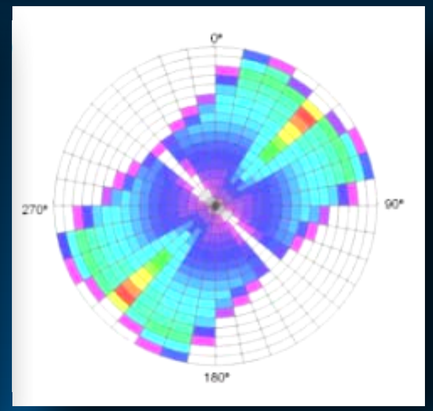
Analysis of Directional Data

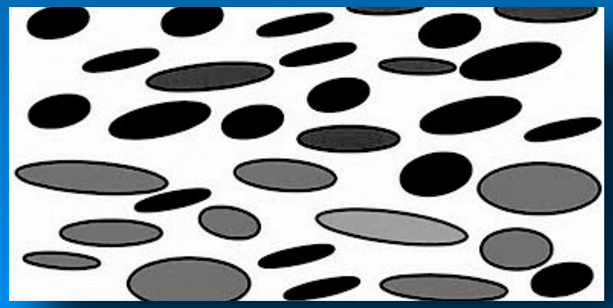
Directional Data

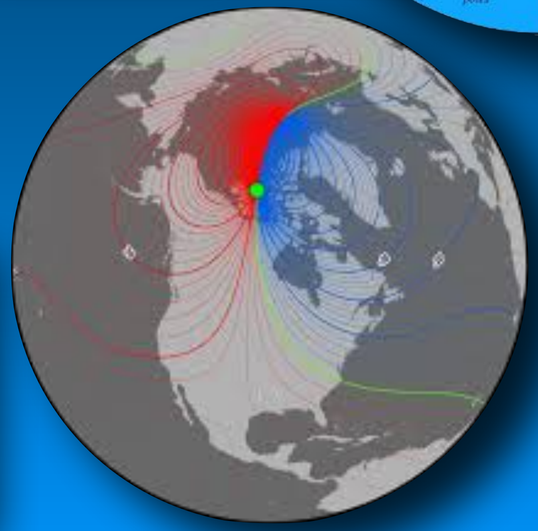
Examples of Directional Data in Earth Science

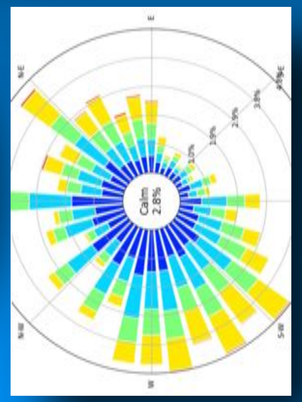


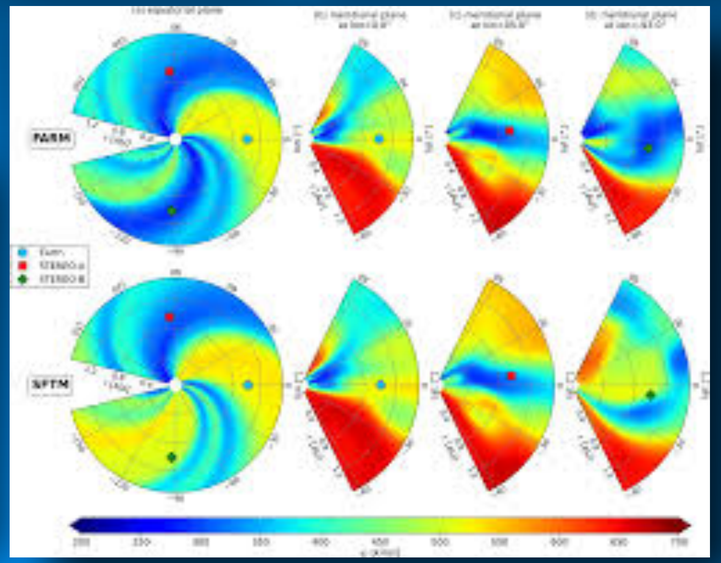






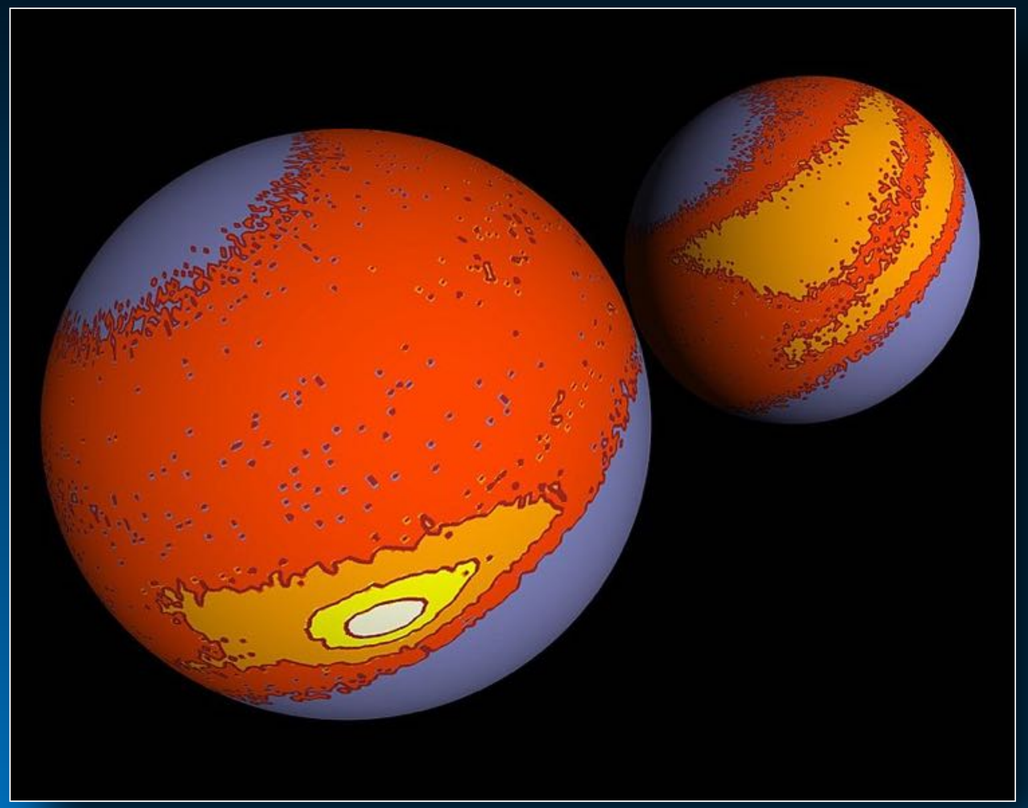






Directional Data

- Data that quantify directions, axes and/or rotations in two or more dimensions.



- Owing to the constrained character of circular, spherical and hyperspherical data, the graphical, data analysis, statistical procedures, and statistical distributions employed in the evaluation of such data are specialized.

- However, directional data analysis has wide application across the earth sciences, extending well into areas not typically thought of as involving time, periodic motions, rotations or angles.

Analysis of Directional Data

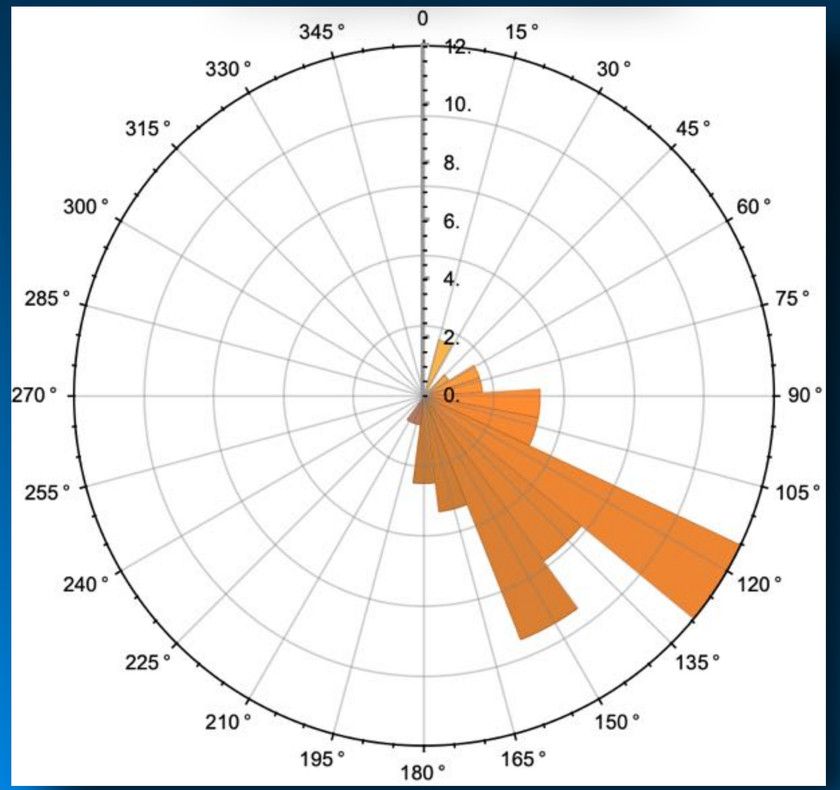
Graphic Presentation of Directional Data

Striation	Angle (°)
1	23
2	27
3	53
4	58
5	64
6	83
7	85
8	88
9	93
10	99
11	100
12	105
13	113
14	113
15	114
16	117
17	121
18	123
19	125
20	126

Striation	Angle (°)
21	129
22	132
23	132
24	132
25	134
26	135
27	137
28	144
29	145
30	145
31	146
32	153
33	155
34	155
35	155
36	157
37	163
38	165
39	171
40	172

Striation	Angle (°)
41	129
42	132
43	132
44	132
45	134
46	135
47	137
48	144
49	145
50	145
51	146

Orientation of Glacial Striations
in Southern Finland



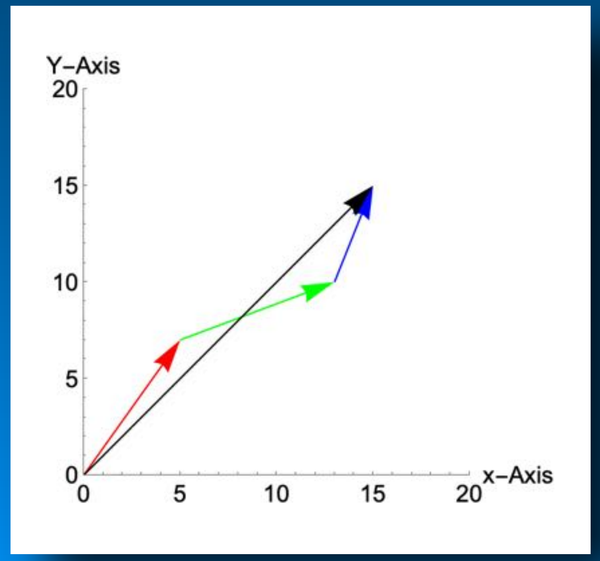
25 Bins

Analysis of Directional Data

Vector Resultant

Data

Vectors	x	y
1	5	7
2	8	3
3	2	5
Resultant	15	15



$$X_r = \sum_{i=1}^{n} x_i$$

$$Y_r = \sum_{i=1}^{n} y_i$$

Where: n = no. of vectors

$$\bar{\theta} = tan^{-1}(Y_r/X_r)$$

$$\bar{d} = \sqrt[2]{Y_r^2 + X_r^2}$$

$X_r = 15$

$Y_r = 15$

$$\bar{\theta} = tan^{-1}(15/15)$$

$$= 0.785 = 45°$$

$$\bar{d} = \sqrt[2]{15^2 + 15^2}$$

$$= 21.213$$

Analysis of Directional Data

To Compare Directions of Samples of Different Sizes

Standardize the Samples

$$\bar{C} = \frac{X_r}{n} \qquad \bar{C} = \frac{1}{n}\sum_{i=1}^{n} cos\theta_i$$

$$\bar{S} = \frac{Y_r}{n} \qquad \bar{S} = \frac{1}{n}\sum_{i=1}^{n} sin\theta_i$$

Resultant

$$R = \sqrt[2]{X_r^2 + Y_r^2}$$

Mean Resultant Length

$$\bar{R} = \frac{R}{n}\sqrt[2]{\bar{C}^2 + \bar{S}^2}$$

Circular Variance

$$s_0^2 = 1 - \bar{R} = (n - R)/n$$

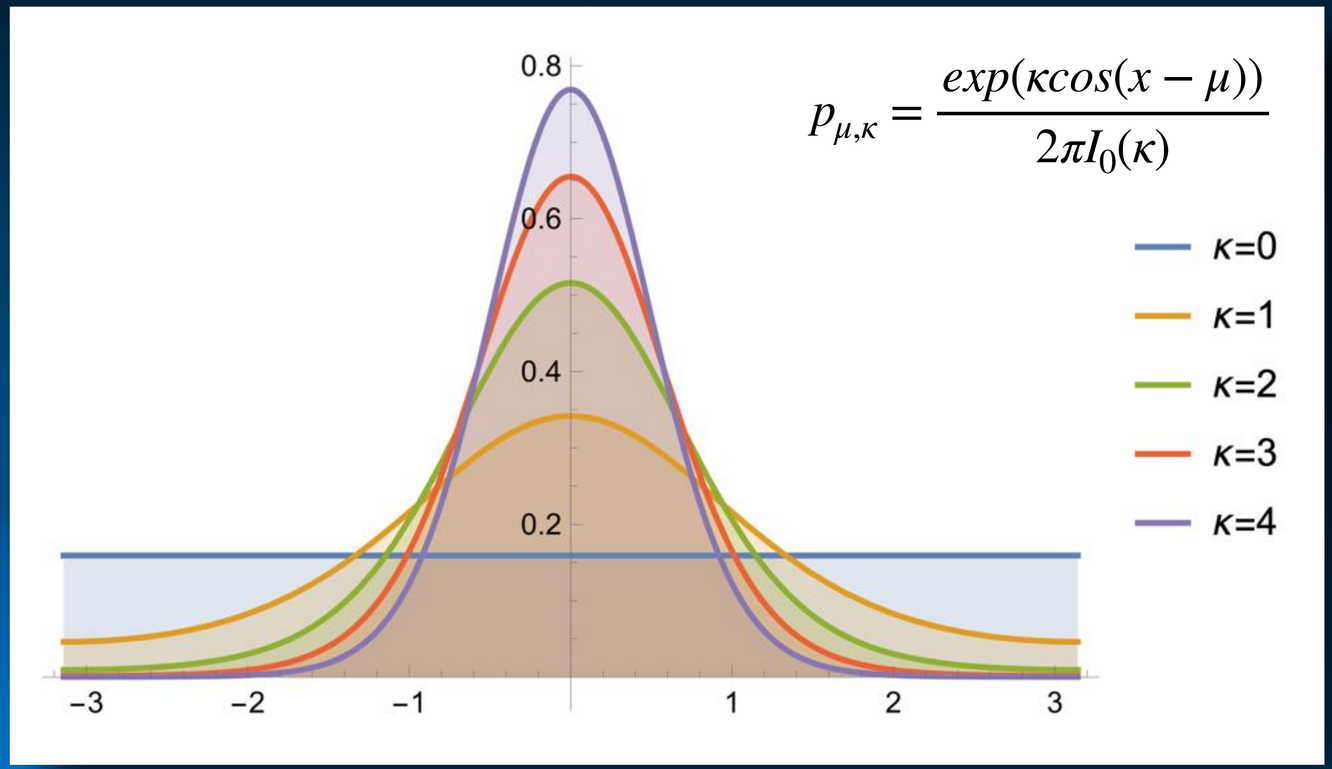
Analysis of Directional Data

von Mises Distribution

Maximum entropy distribution for circular data – also referred to as the circular normal distribution. This distribution is referenced to the sample mean angle (μ) and a measure of concentration (κ) which is analogous to the sample variance.

$\kappa = 0$: all angles are equally probable;

$\kappa > 2$: the distribution becomes highly concentrated about the angle μ.



The von Mises distribution is used extensively in tests designed to determine whether a single angle, or a group of angles differs significantly from a limiting condition.

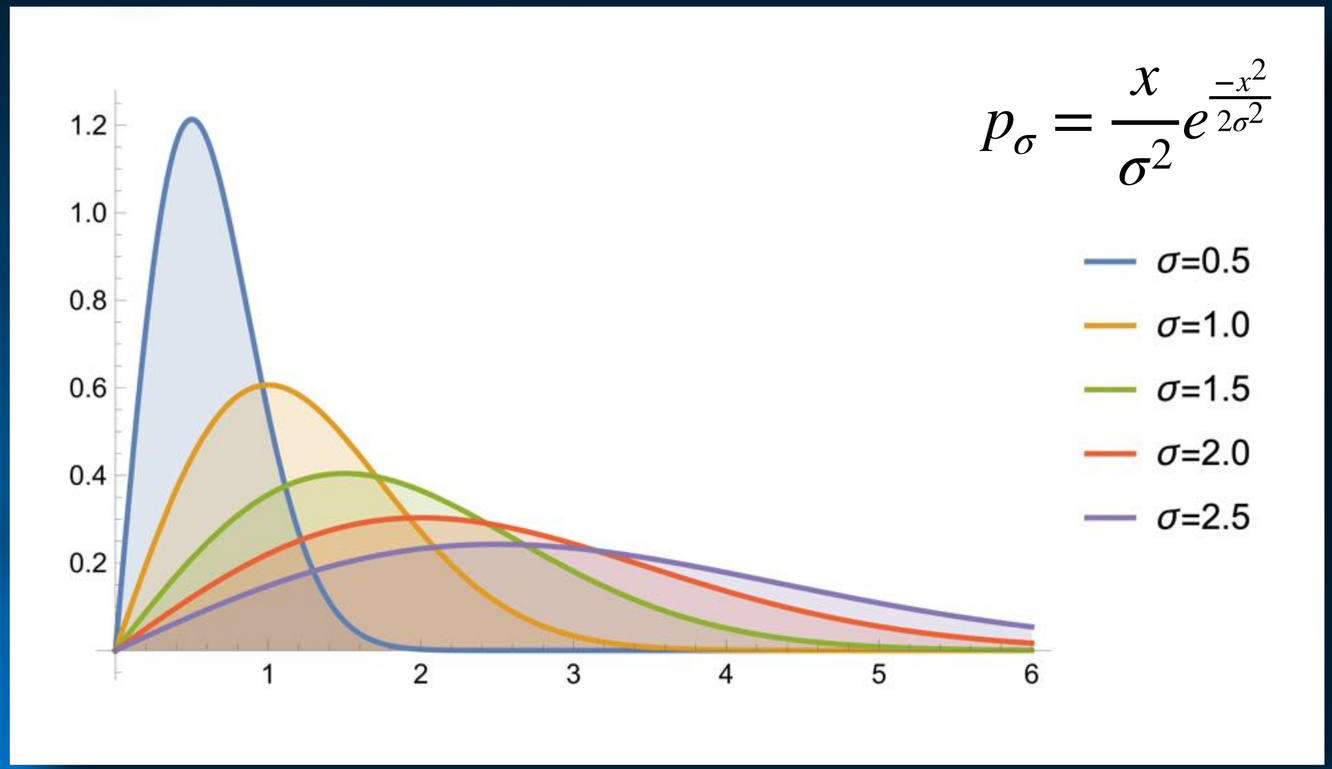
Maximum Likelihood Estimates of von Mises κ

Mean R	Kappa	Mean R	Kappa	Mean R	Kappa	Mean R	Kappa
0.00	0.00000	0.26	0.53863	0.52	1.22350	0.78	2.64613
0.01	0.02000	0.27	0.56097	0.53	1.25672	0.79	2.75382
0.02	0.04001	0.28	0.58350	0.54	1.29077	0.80	2.87129
0.03	0.06003	0.29	0.60625	0.55	1.32570	0.81	3.30020
0.04	0.08006	0.30	0.62922	0.56	1.36156	0.82	3.14262
0.05	0.10013	0.31	0.65242	0.57	1.39842	0.83	3.30114
0.06	0.12022	0.32	0.67587	0.58	1.43635	0.84	3.47901
0.07	0.14034	0.33	0.69958	0.59	1.47543	0.85	3.68401
0.08	0.16051	0.34	0.72356	0.60	1.51574	0.86	3.91072
0.09	0.18073	0.35	0.74783	0.61	1.55738	0.87	4.17703
0.10	0.20101	0.36	0.77241	0.62	1.60044	0.88	4.48876
0.11	0.22134	0.37	0.79730	0.63	1.64506	0.89	4.85871
0.12	0.24175	0.38	0.82253	0.64	1.69134	0.90	5.30470
0.13	0.26223	0.39	0.84812	0.65	1.73945	0.91	5.85220
0.14	0.28279	0.40	0.87408	0.66	1.78953	0.92	6.53940
0.15	0.30344	0.41	0.90043	0.67	1.84177	0.93	7.42570
0.16	0.32419	0.42	0.92720	0.68	1.89637	0.94	8.61040
0.17	0.34503	0.43	0.95440	0.69	1.95357	0.95	10.27160
0.18	0.36599	0.44	0.98207	0.70	2.01363	0.96	12.76610
0.19	0.38707	0.45	1.01022	0.71	2.07685	0.97	16.92660
0.20	0.40828	0.46	1.03889	0.72	2.14359	0.98	25.25220
0.21	0.42962	0.47	1.06810	0.73	2.21425	0.99	50.24210
0.22	0.45110	0.48	1.09788	0.74	2.28930	1.00	∞
0.23	0.47273	0.49	1.12828	0.75	2.36930		
0.24	0.49453	0.50	1.15932	0.76	2.45490		
0.25	0.51649	0.51	1.19105	0.77	2.54686		

Analysis of Directional Data

Rayleigh Distribution

A continuous probability distribution for non-negative random variables, similar to the χ^2 distribution with 2 degrees of freedom. Used in vector Tests where the magnitude of a vector in a 2D place is related to Its directional components. This Distribution assumes the distribution of magnitudes and directions are uncorrelated, normally distributed, have equal variances and 0.0 mean.



Owing to this highly limiting, and somewhat unrealistic, conditions the Rayleigh distribution is often replaced by the Weibull distribution in many practical applications. Regardless, this distribution is used in a number of simple, directional statistics tests.

Analysis of Directional Data

Rayleigh Distribution

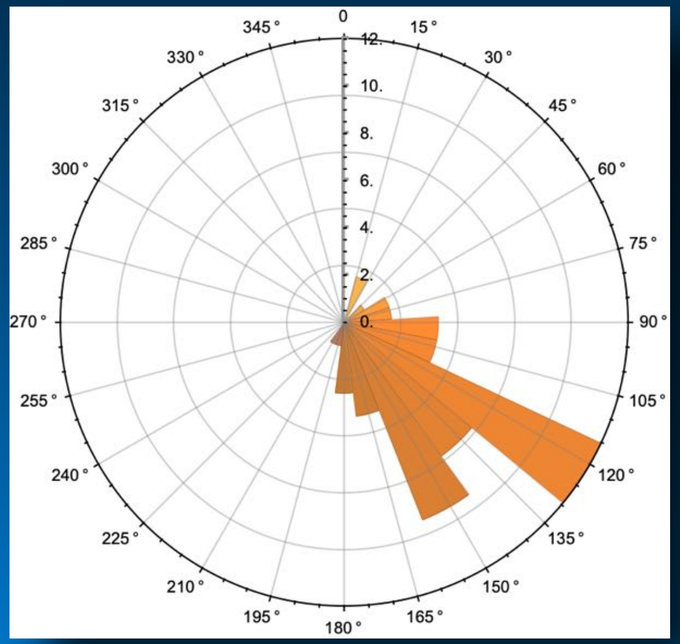
Critical z Values for the Rayleigh's Test
Taken from Zar, 1981 Table B.32

n	α: 0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
6	0.734	1.639	2.274	2.865	3.576	4.058	4.491	4.985	5.297
7	0.727	1.634	2.278	2.885	3.627	4.143	4.617	5.181	5.556
8	0.723	1.631	2.281	2.899	3.665	4.205	4.710	5.322	5.743
9	0.719	1.628	2.283	2.910	3.694	4.252	4.780	5.430	5.885
10	0.717	1.626	2.285	2.919	3.716	4.289	4.835	5.514	5.996
11	0.715	1.625	2.287	2.926	3.735	4.319	4.879	5.582	6.085
12	0.713	1.623	2.288	2.932	3.750	4.344	4.916	5.638	6.158
13	0.711	1.622	2.289	2.937	3.763	4.365	4.947	5.685	6.219
14	0.710	1.621	2.290	2.941	3.774	4.383	4.973	5.725	6.271
15	0.709	1.620	2.291	2.945	3.784	4.398	4.996	5.759	6.316
16	0.708	1.620	2.292	2.948	3.792	4.412	5.015	5.789	6.354
17	0.707	1.619	2.292	2.951	3.799	4.423	5.033	5.815	6.388
18	0.706	1.619	2.293	2.954	3.806	4.434	5.048	5.838	6.418
19	0.705	1.618	2.293	2.956	3.811	4.443	5.061	5.858	6.445
20	0.705	1.618	2.294	2.958	3.816	4.451	5.074	5.877	6.469
21	0.704	1.617	2.294	2.960	3.821	4.459	5.085	5.893	6.491
22	0.704	1.617	2.295	2.961	3.825	4.466	5.095	5.908	6.510
23	0.703	1.616	2.295	2.963	3.829	4.472	5.104	5.922	6.528
24	0.703	1.616	2.295	2.964	3.833	4.478	5.112	5.935	6.544
25	0.702	1.616	2.296	2.966	3.836	4.483	5.120	5.946	6.559

Analysis of Directional Data

Rayleigh's Test for a Preferred Trend

Orientation of Glacial Striations in Southern Finland



$$X_r = \sum_{i=1}^{n} cos\ \theta_i \qquad X_r = -25.7933$$

$$Y_r = \sum_{i=1}^{n} sin\ \theta_i \qquad Y_r = 31.6367$$

$$R = \sqrt[2]{X_r^2 + Y_r^2} \qquad R = 40.8188$$

$$\bar{R} = R/n \qquad \bar{R} = 0.8004$$

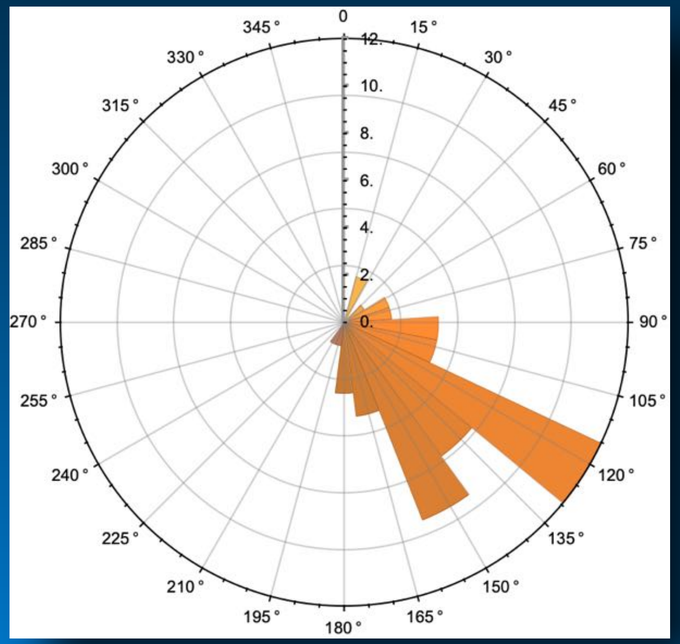
$$\bar{R}_{\alpha=0.05,n=51} = 0.2981$$

Reject H_0 at a = 0.05. The data exhibit a preferred orientation.

Analysis of Directional Data

Rayleigh's Test for Comparison to a Specific Trend

Orientation of Glacial Striations in Southern Finland



This test can be made exact, but this requires the use of extensive charts to enable setting of the correct critical value (Stephens, 1969). An alternative, approximate approach is to calculate the confidence interval on the mean vector direction and use that as a standard to which any specific trend might be compared.

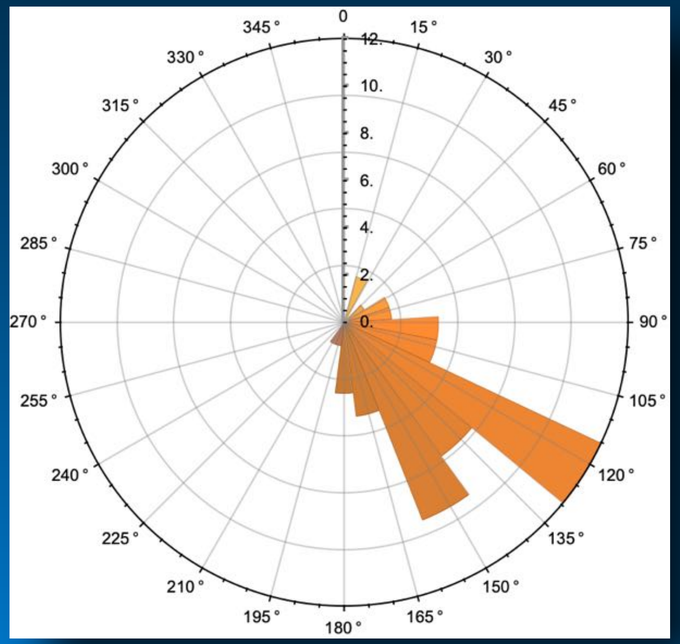
Mean Vector: $\bar{\theta} = tan^{-1}(Y_r/X_r)$

Standard Error: $s_e = \frac{1}{\sqrt[2]{n\bar{R}\kappa}}$

Analysis of Directional Data

Rayleigh's Confidence-Interval Test for Comparison to a Specific Trend

Orientation of Glacial Striations in Southern Finland



$$\bar{\theta} = tan^{-1}(Y_r/X_r) \qquad \bar{\theta} = 129.19°$$

$$s_e = \frac{1}{\sqrt[2]{n\bar{R}\kappa}} \qquad s_e = 5.2924°$$

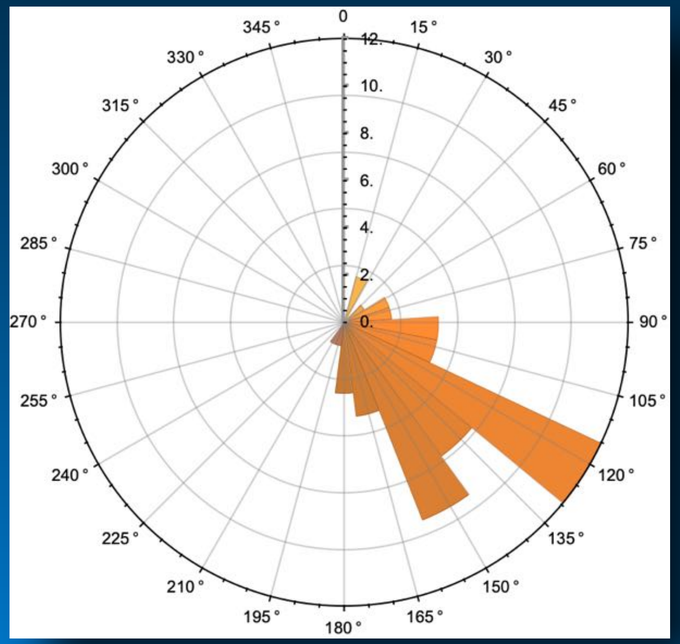
$$CI_{\alpha=0.05} = Z_\alpha s_e$$

$$CI_{\alpha=0.05} = 123.898° - 134.483°$$

Analysis of Directional Data

V Test for Comparison to a Specific Trend

Orientation of Glacial Striations in Southern Finland



$$V = \frac{1}{n}\sum_{i=0}^{n} cos(\theta_i - \mu_0)$$

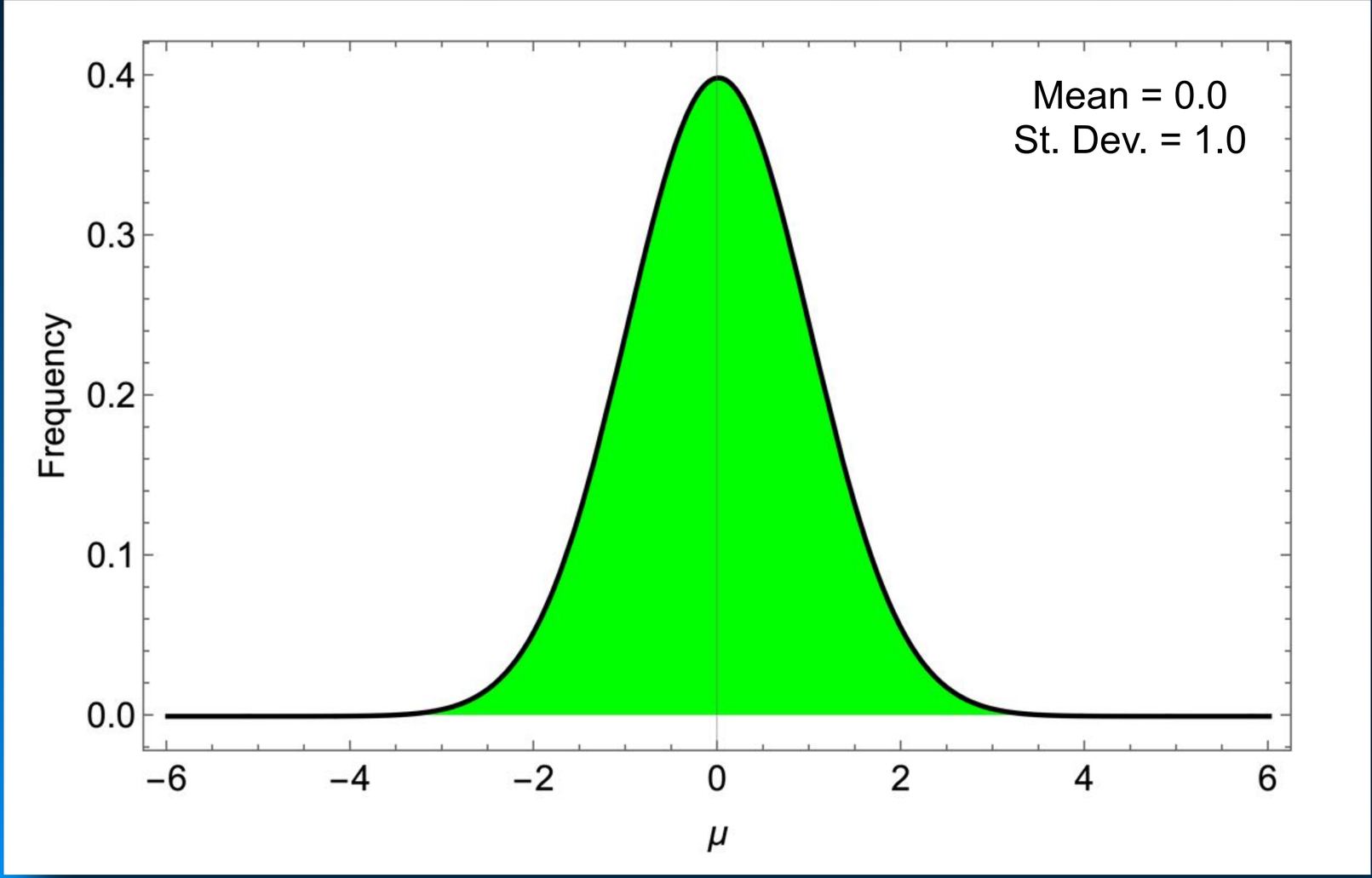
Where: μ_0 - specific trend angle.

$$\bar{R} = \sqrt[2]{(\frac{1}{n}\sum_{i=1}^{n} cos(\theta_i)) + (\frac{1}{n}\sum_{i=1}^{n} sin(\theta_i))}$$

$$u = \bar{R} \cdot \sqrt[2]{2n} \cdot cos(\bar{\theta} - \mu_0)$$

Analysis of Directional Data

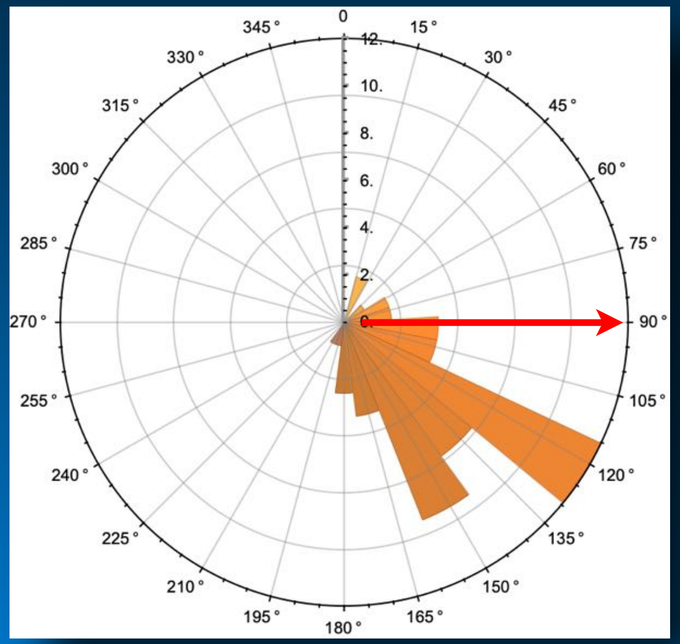
V Test is Referenced to the Standard Normal Distribution



Analysis of Directional Data

V Test for Comparison to a Specific Trend: Example

Orientation of Glacial Striations in Southern Finland



$$\mu_0 = 90°$$

$$V = \frac{1}{n}\sum_{i=0}^{n} cos(\theta_i - \mu_0) \qquad V = 0.6203$$

$$\bar{R} = \sqrt[2]{(\frac{1}{n}\sum_{i=1}^{n} cos(\theta_i)) + (\frac{1}{n}\sum_{i=1}^{n} sin(\theta_i))} \qquad \bar{R} = 0.3385$$

$$u = \bar{R} \cdot \sqrt[2]{2n} \cdot cos(\bar{\theta} - \mu_0) \qquad u = 2.7761$$

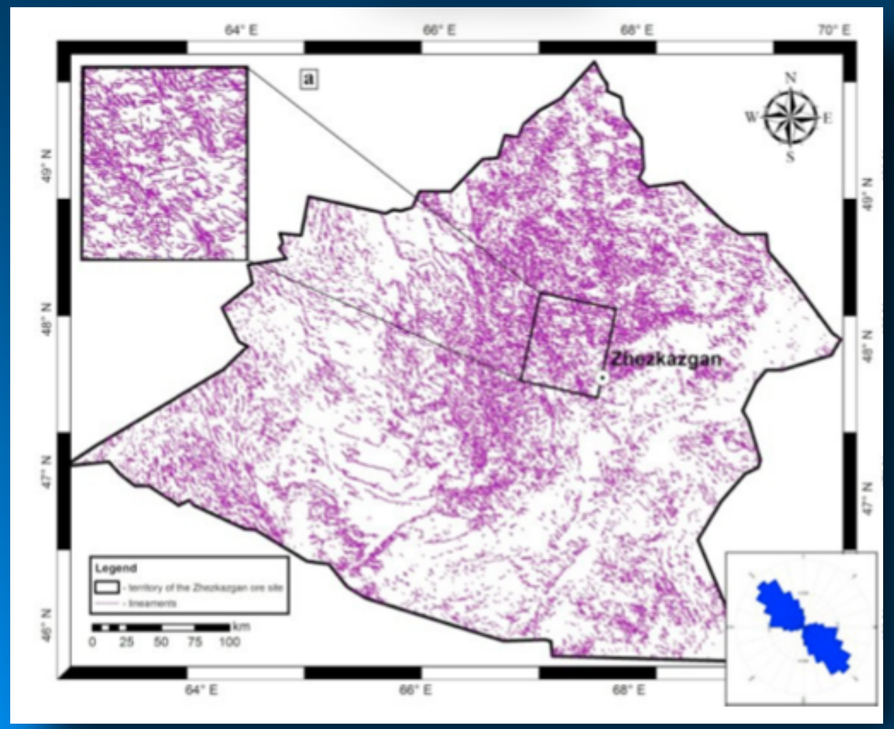
$$\alpha = 0.9967$$

Analysis of Directional Data

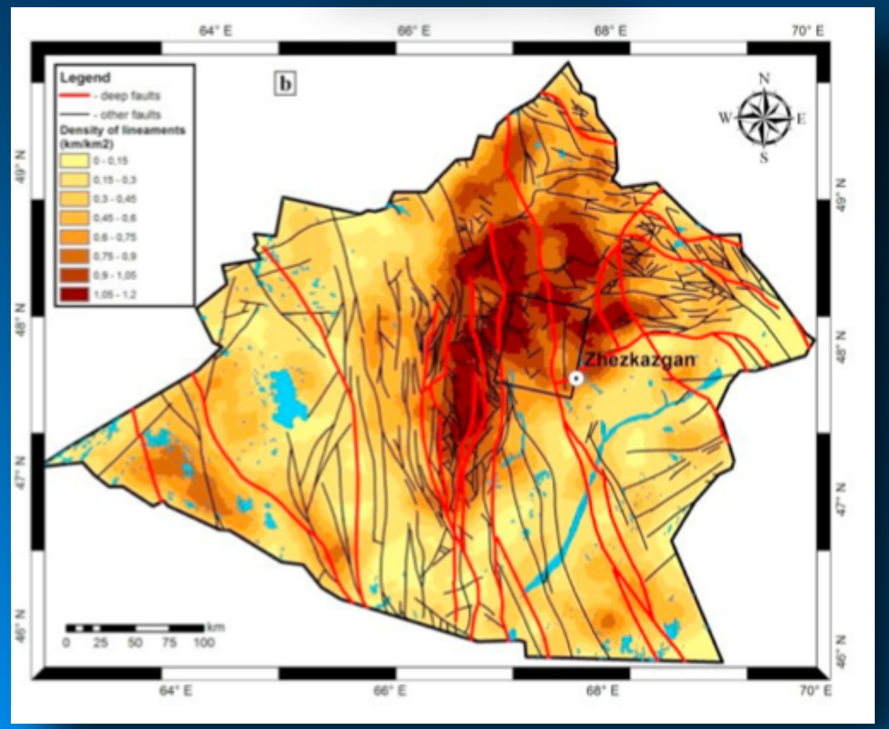
Directional Data: Example Analysis

Landsat satellite images commonly exhibit large-scale linear structuring of the Earth's surface. But the origin of these lineations is unknown. These two Bangladesh datasets can be used to test for an association between Landsat-identified lineaments and the axes of subsurface anticlines.

Linements



Anticline Axes



Zhumabek et al. (2017)

Analysis of Directional Data

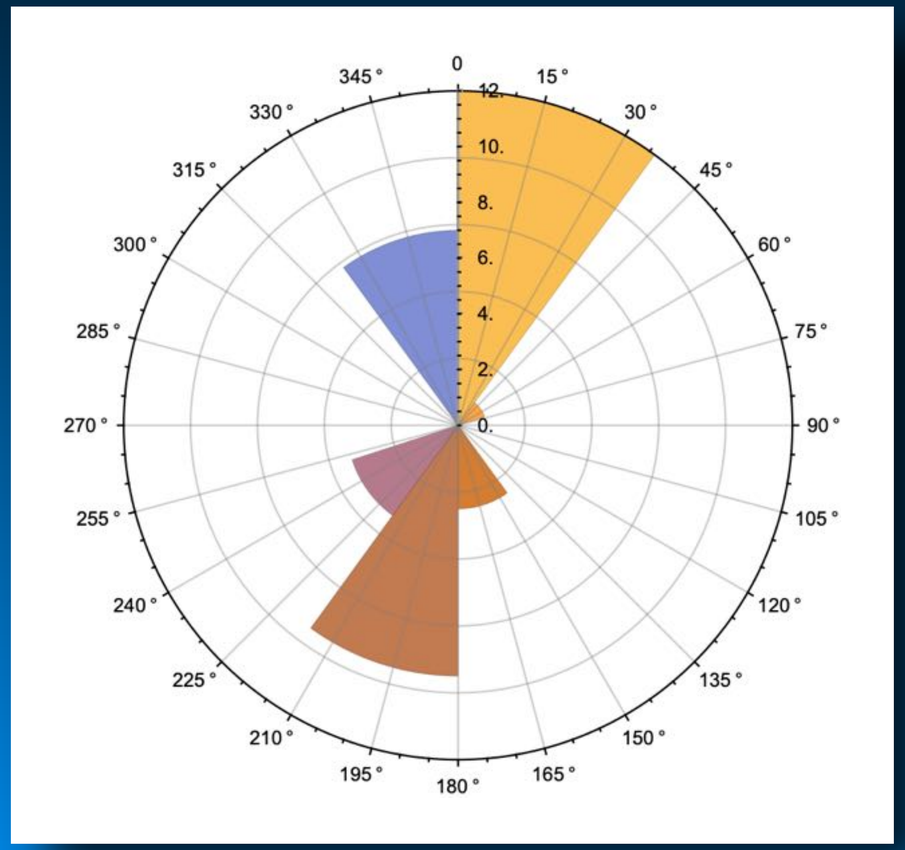
Test for Equality of Samples in Directional Data

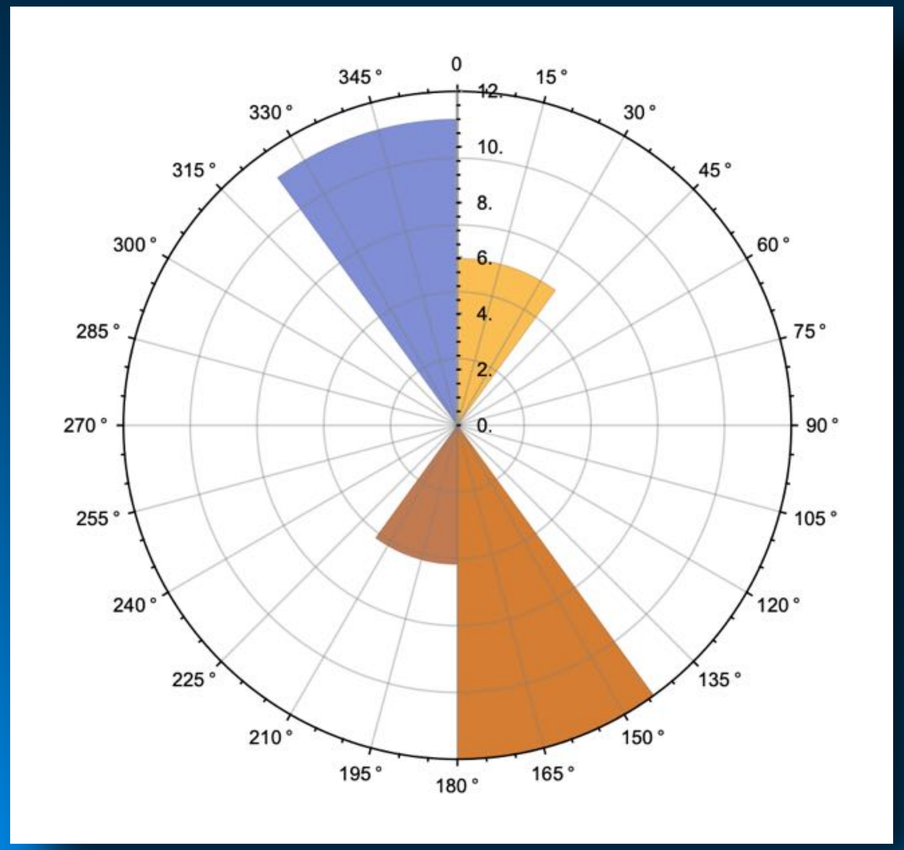
LandSat Lineaments (n=36)			
350	32	15	8
214	192	16	26
356	218	198	221
350	18	221	342
160	205	35	337
2	171	196	14
184	246	175	25
354	213	26	212
42	354	13	202

Anticline Axes (n=34)			
12	16	14	5
192	202	169	163
186	186	24	344
343	346	161	341
339	150	169	336
351	156	159	352
152	162	341	181
348	158	156	
330	162	20	

Analysis of Directional Data

Test for Equality of Samples in Directional Data





Analysis of Directional Data

Test for Equality of Samples in Directional Data

The equality of two mean vectors can be tested by comparing the resultant vectors of the two groups added together to the resultant vector of the pooled data. If the two samples were drawn from the same population this angular difference should be negligible (H_0), but if they were drawn from different populations it should be large (H_1).

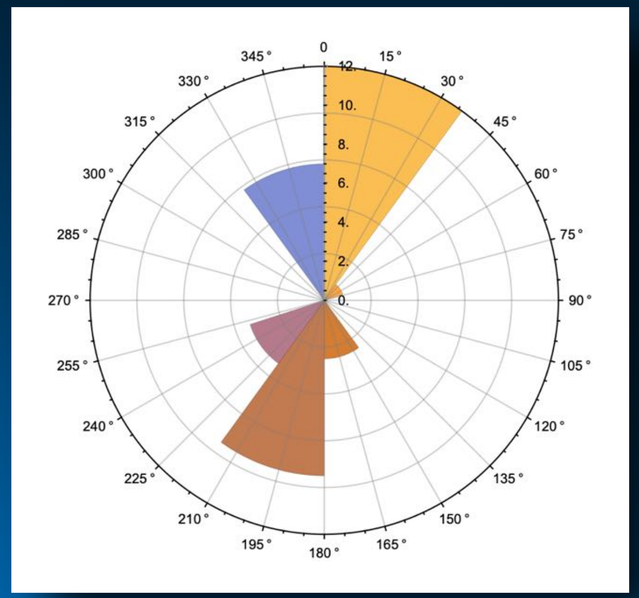
$$If\ \kappa > 10 \quad F_{1,n-2} = \frac{(n-2)\cdot(R_1 + R_2 - R_{pool})}{(n - R_1 - R_2)}$$

$$If\ \kappa \leq 10 \quad F_{1,n-2} = (1 + \frac{3}{8k})) \cdot (\frac{(n-2)\cdot(R_1 + R_2 - R_{pool})}{(n - R_1 - R_2)})$$

$$where: n = pooled\ sample\ size$$
$$\kappa = max.\ likelihood\ of\ \overline{R}\ concentration\ factor$$

Analysis of Directional Data

Test for Equality of Samples in Directional Data



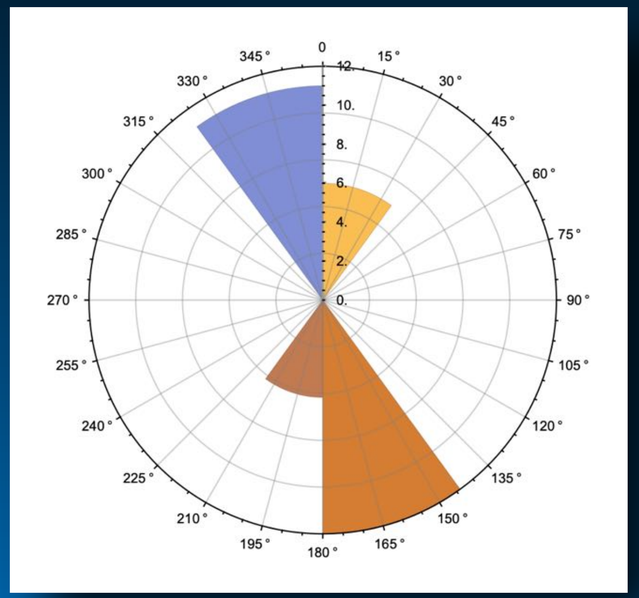
$$R_{antcline} = 14.6873$$

$$R_{lineation} = 18.19$$

$$R_{pooled} = 22.289$$

$$\bar{R}_{pooled} = 0.3184$$

$$\kappa = 0.67580$$



$$F = (1 + \frac{1}{3 * 0.6758})(\frac{(70 - 2)(14.6873 + 18.19 - 22.289)}{(70 - 14.6873 - 18.19))}$$

$$F = (1.5549)(19.3953) \qquad F = 30.1577 \qquad F_{\alpha=0.05,dof=68} = 3.99 \qquad \text{Reject } H_0$$

Spatial & Directional Data Analysis

Prof. Norman MacLeod
School of Earth Sciences & Engineering, Nanjing University