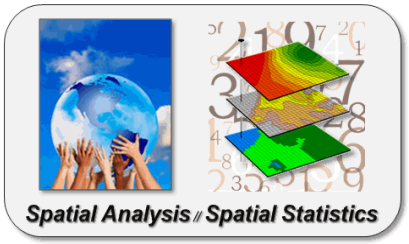
*Spatial*STEM:

A Mathematical/Statistical Framework for Understanding and Communicating Map Analysis and Modeling



Part 3) **Spatial Statistics**. *Spatial Statistics* involves quantitative analysis of the "**numerical context**" of mapped data, such as characterizing the geographic distribution, relative comparisons, map similarity or correlation within and among data layers. Spatial Analysis and Spatial Statistics form a map-*ematics* that uses **sequential processing** of analytical operators to develop complex map analyses and models. Its approach is similar to traditional statistics except the variables are entire sets of geo-registered mapped data.

This PowerPoint with notes and online links to further reading is posted at

www.innovativegis.com/basis/Workshops/NGA2015/

Presented by

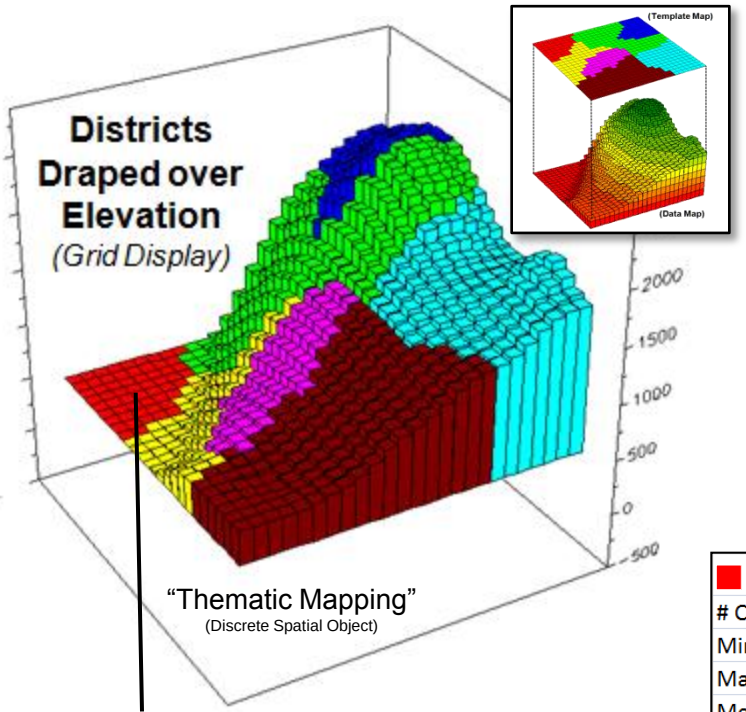
Joseph K. Berry

Adjunct Faculty in Geosciences, Department of Geography, University of Denver
Adjunct Faculty in Natural Resources, Warner College of Natural Resources, Colorado State University
Principal, Berry & Associates // Spatial Information Systems

Email: jberry@innovativegis.com — Website: www.innovativegis.com/basis

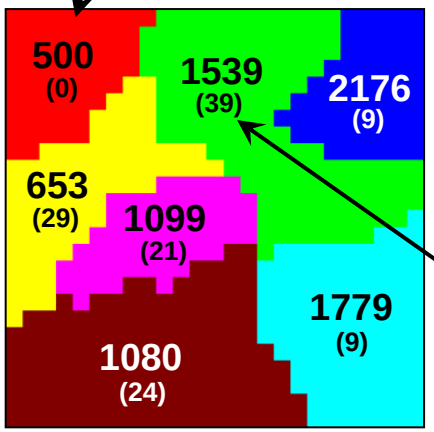
Thematic Mapping ≠ Map Analysis *(Average elevation by district)*

Thematic Mapping assigns a "typical value" to irregular geographic "puzzle pieces" (map features) describing the characteristics/condition without regard to their **continuous spatial distribution** ***(non-quantitative characterization)***



"Thematic Mapping"
(Discrete Spatial Object)

Average Elevation
of Districts



… at least include **CoffVar** in Thematic Mapping results

…**average** is assumed to be everywhere the same within each puzzle piece (± 1 Stdev)

District #1	
# Cells	58.0
Min	500.0
Max	500.0
Median	500.0
Average	500.0
StDev	0.0
CoffVar	0.0

Best

District #2	
# Cells	135.0
Min	500.0
Max	2499.0
Median	1499.0
Average	1538.9
StDev	597.5
CoffVar	38.8

Worst

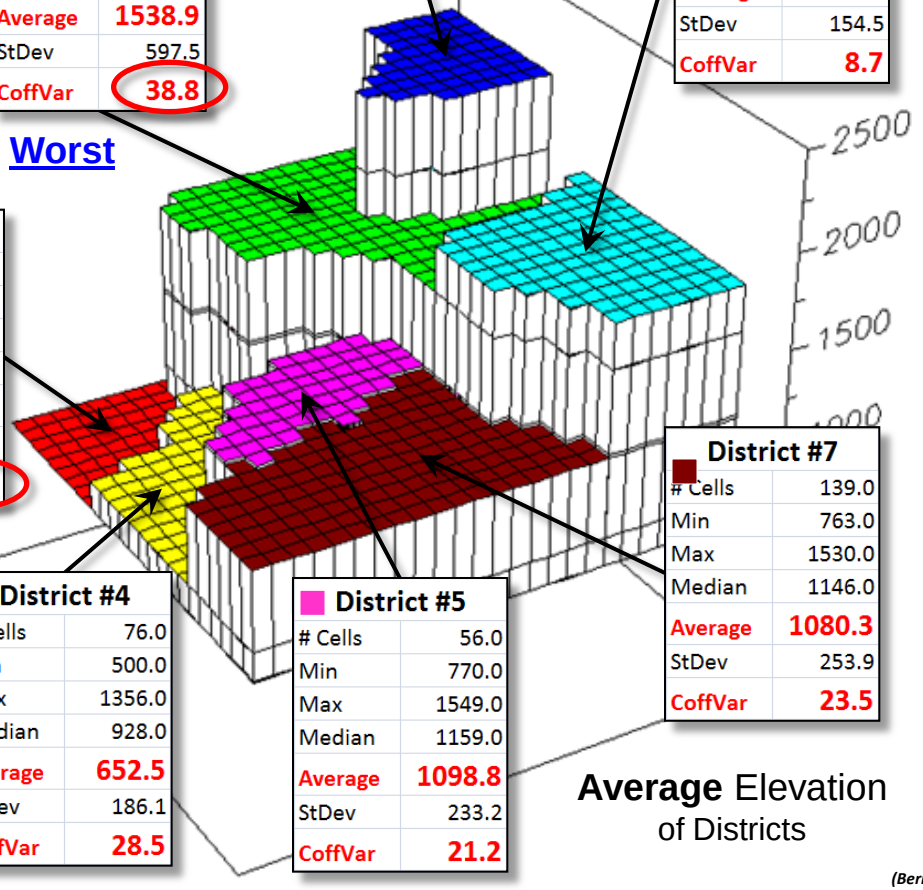
District #3	
# Cells	59.0
Min	1786.0
Max	2500.0
Median	2143.0
Average	2176.0
StDev	185.3
CoffVar	8.5

District #4	
# Cells	76.0
Min	500.0
Max	1356.0
Median	928.0
Average	652.5
StDev	186.1
CoffVar	28.5

District #5	
# Cells	56.0
Min	770.0
Max	1549.0
Median	1159.0
Average	1098.8
StDev	233.2
CoffVar	21.2

District #6	
# Cells	102.0
Min	1507.0
Max	2152.0
Median	1829.0
Average	1778.9
StDev	154.5
CoffVar	8.7

District #7	
# Cells	139.0
Min	763.0
Max	1530.0
Median	1146.0
Average	1080.3
StDev	253.9
CoffVar	23.5



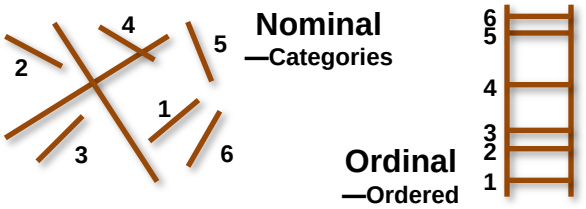
Average Elevation
of Districts

(Berry)

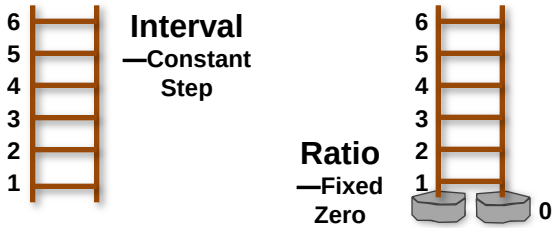
Spatial Data Perspectives *(numerically defining the What in "Where is What")*

Numerical Data Perspective: *how numbers are distributed in "Number Space"*

- ***Qualitative**: deals with unmeasurable qualities (very few math/stat operations available)*
  - – ***Nominal numbers*** are independent of each other and do not imply ordering – like scattered pieces of wood on the ground
  - – ***Ordinal numbers*** imply a definite ordering from small to large – like a ladder, however with varying spaces between rungs



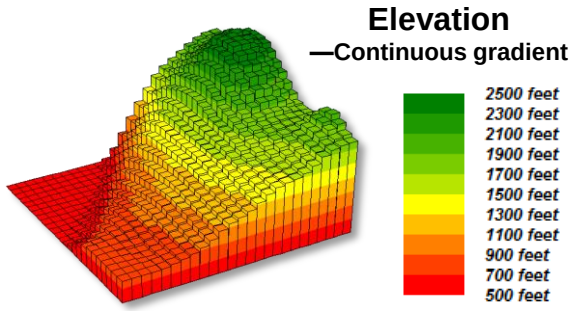
- ***Quantitative**: deals with measurable quantities (a wealth of math/stat operations available)*
  - – ***Interval numbers*** have a definite ordering and a constant step – like a typical ladder with consistent spacing between rungs
  - – ***Ratio numbers*** has all the properties of interval numbers plus a clear/constant definition of 0.0 – like a ladder with a fixed base.

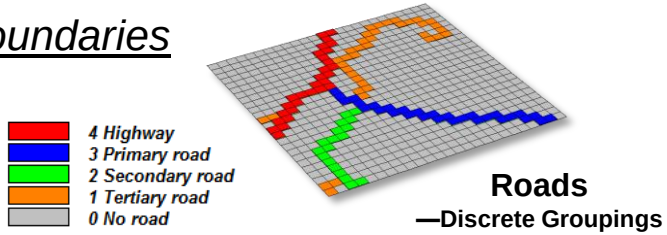


- ***Binary**: a special type of number where the range is constrained to just two states— such as 1=forested, 0=non-forested*

Spatial Data Perspective: *how numbers are distributed in "Geographic Space"*

- ***Choropleth* numbers** *form sharp/unpredictable boundaries in geographic space – e.g., a road "map"*





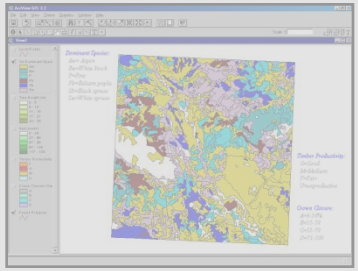
- ***Isopleth* numbers** *form continuous and often predictable gradients in geographic space – e.g., an elevation "surface"*

(Berry)

Overview of Map Analysis Approaches

(Spatial Analysis and Spatial Statistics)

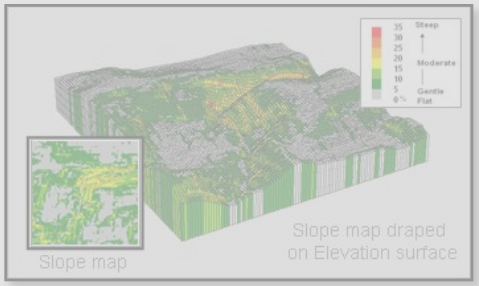
Traditional GIS



Forest Inventory Map

- Points, Lines, Polygons
- Discrete Objects
- Mapping and Geo-query

Spatial Analysis

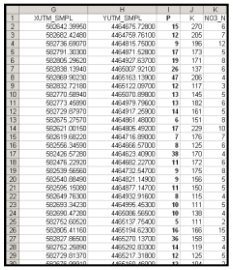


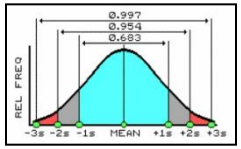
Elevation (Surface)

- Cells, Surfaces
- Continuous Geographic Space
- Contextual Spatial Relationships

…last session

Traditional Statistics

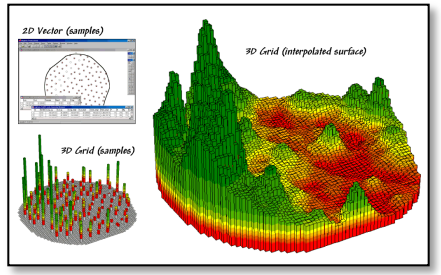




Minimum= 5.4 ppm
Maximum= 103.0 ppm
Mean= 22.4 ppm
StDEV= 15.5

- **Mean, StDev (Normal Curve)**
- **Central Tendency**
- **Typical Response (scalar)**

Spatial Statistics



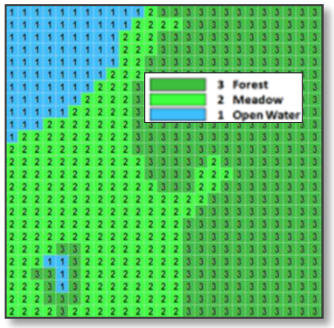
Spatial Distribution (Surface)

- **Map of Variance (gradient)**
- **Spatial Distribution**
- **Numerical Spatial Relationships**

(Berry)

Spatial Statistics Operations *(Numerical Context)*

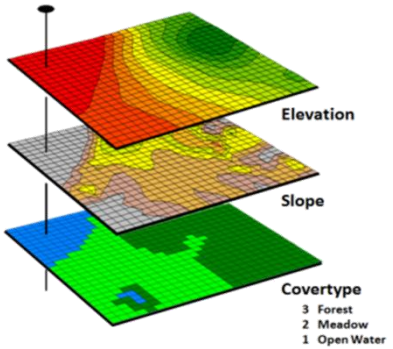
GIS as "Technical Tool" *(Where is What)* vs. "**Analytical Tool**" *(Why, So What and What if)*



Grid Layer

Spatial Statistics

Map Stack



Spatial Statistics seeks to map the spatial variation in a data set instead of focusing on a single typical response (central tendency) ignoring the data's spatial distribution/pattern, and thereby provides a mathematical/statistical framework for *analyzing* and *modeling* the

Numerical Spatial Relationships

within and among grid map layers

Map Analysis Toolbox

Statistical Perspective:

…let's consider some examples →

Basic Descriptive Statistics *(Min, Max, Median, Mean, StDev, etc.)*

Basic Classification *(Reclassify, Contouring, Normalization)*

Map Comparison *(Joint Coincidence, Statistical Tests)*

✓ ***Unique Map Statistics*** *(Roving Window and Regional Summaries)*

✓ ***Surface Modeling*** *(Density Analysis, Spatial Interpolation)*

Advanced Classification *(Map Similarity, Maximum Likelihood, Clustering)*

Predictive Statistics *(Map Correlation/Regression, Data Mining Engines)*

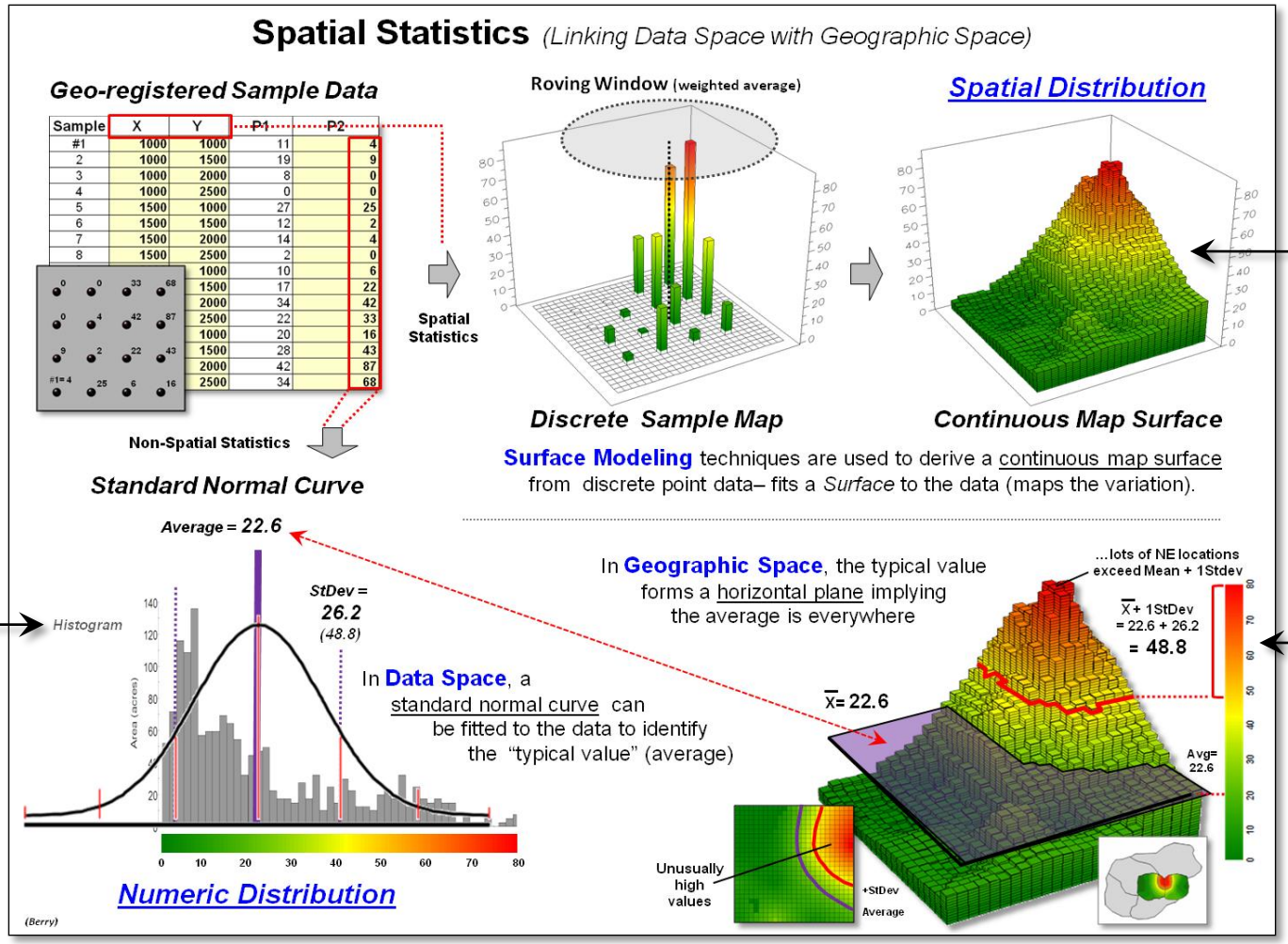
(Berry)

Spatial Statistics *(Linking Data Space with Geographic Space …review from "**Future Directions**" seminar)*

Traditional Statistics fits a Standard Normal Curve (2D density function) to identify the **typical value** in a data set (*Mean*) and its typical variation (*Standard Deviation*) in abstract data space.

Surface Modeling uses Spatial Interpolation to fit a continuous surface (3D density function) that maps the **spatial distribution** (variation) in geographic space. *Spatial Analyst* IDW, Kriging, Spline and Natural Neighbors commands.

Unusually High Locations are identified as locations greater than the *Mean* plus one *Standard Deviation* (**upper tail**). *Spatial Analyst* Reclass command.



Video of "***Future Directions in Map Analysis and Modeling***" seminar can be viewed at https://www.youtube.com/watch?v=YA-mGpc20vc

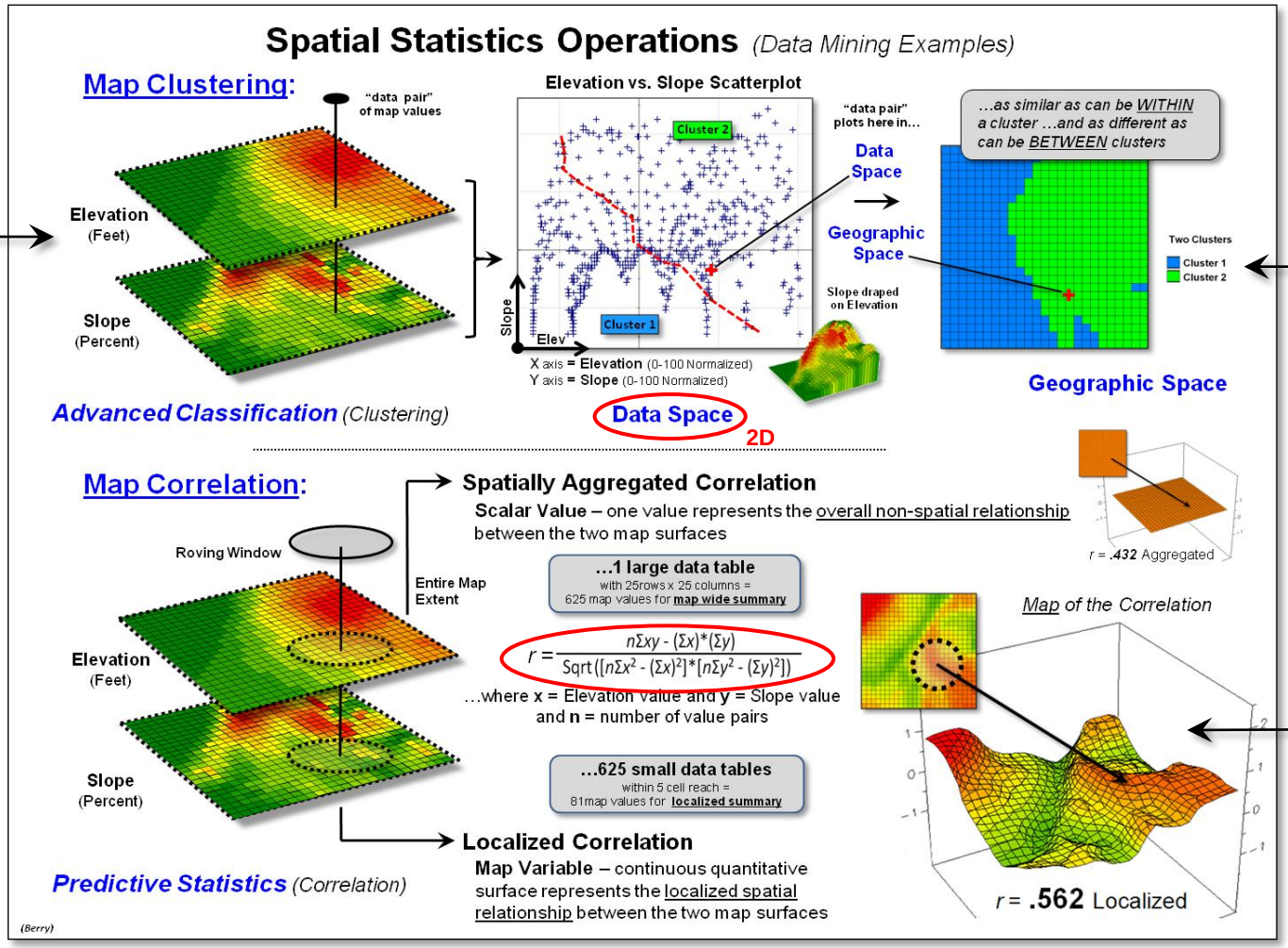
(Berry)

Spatial Statistics Operations *(Data Mining Examples …review from "**Future Directions**" seminar)*

Data Space plots pairs of spatially coincident data values (2D scatter plot) in abstract data space to identify **data pairing relationships** (e.g., low-low, …, high-high) but can be expanded to tuples in n-dimensional space.

Clustering uses the Pythagorean Theorem in calculating **Data Distance** between data parings to quantitatively establish Clusters (groupings with minimal inter-cluster distances). *Spatial Analyst* Iso Cluster command.

A **Correlation Map** is generated by repetitive evaluation of the Correlation Equation within a **Roving Window** of nearby data pairings. *Spatial Analyst* dropped Correlation AML command.



Video of "***Future Directions in Map Analysis and Modeling***" seminar can be viewed at https://www.youtube.com/watch?v=YA-mGpc20vc

(Berry)

Spatial Variable Dependence *(the keystone concept in Spatial Statistics)*

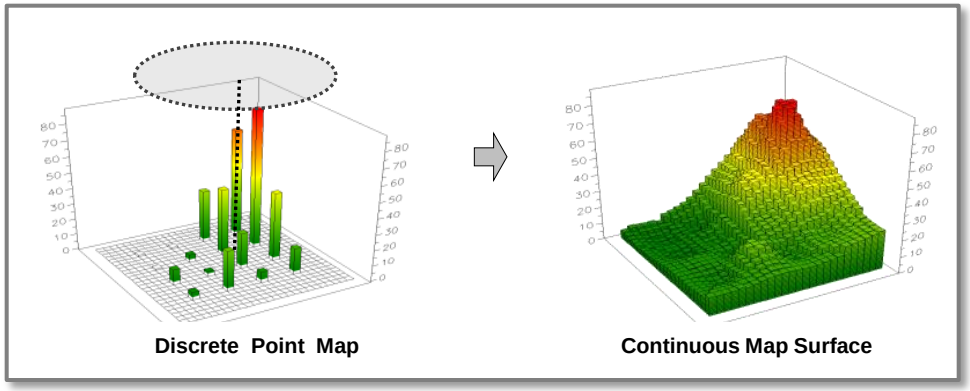
*There are **two types** of spatial dependency based on*

*…"what occurs at a location in geographic space is **related to**" —*

*1) …the **conditions of that variable at nearby locations**, termed*

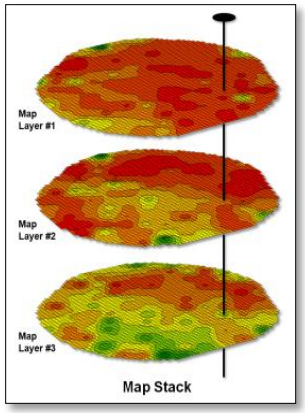
Spatial Autocorrelation *(intra-variable dependence; within a map layer)*

Surface Modeling **–** *identifies the continuous spatial distribution implied in a set of discrete point samples*



*2) …the **conditions of other variables at that location**, termed*

Spatial Correlation *(inter-variable dependence; among map layers)*



Spatial Data Mining **–** *investigates spatial relationships among multiple map layers by spatially evaluating traditional statistical procedures*

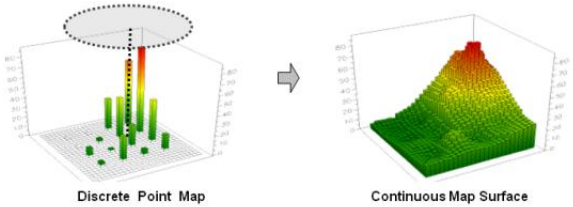
Map Stack *– relationships among maps are investigated by aligning grid maps with a common configuration— same **#cols/rows**, **cell size** and **geo-reference***

Data Shishkebab *– within a statistical context, each map layer represents a **Variable**; each grid space a **Case**; and each value a **Measurement** …with all of the rights, privileges, and responsibilities of non-spatial mathematical, numerical and statistical analysis*

(Berry)

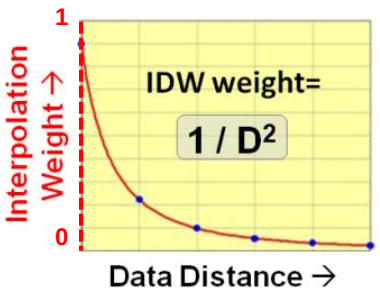
Surface Modeling Approaches

…spatial dependency within a single map layer (Spatial Autocorrelation)



Surface Modeling identifies the continuous spatial distribution implied in a set of discrete point data using one of **four basic approaches**—

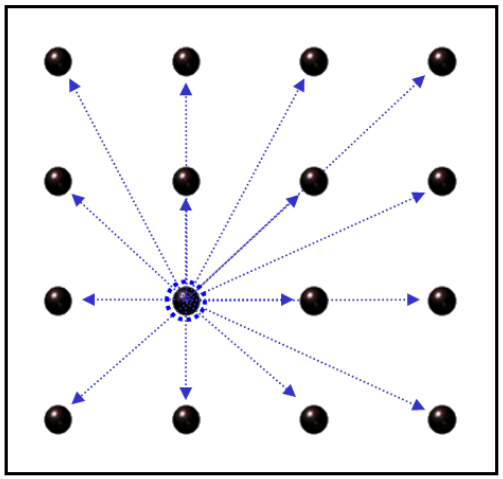
- **Map Generalization** "best fits" a **polynomial equation** to the entire set of geo-registered data values
- **Geometric Facets** "best fits" a set of **geometric shapes** (e.g., irregularly sized/shaped triangles) to the data values
- **Density Analysis** "counts or sums" data values occurring within a **roving window** (Qualitative/Quantitative)
- **Spatial Interpolation** "weight-averages" data values within a **roving window** based on a mathematical relationship relating *Data Variation* to *Data Distance* that assumes "*nearby things are more alike than distant things*" (Quantitative)…

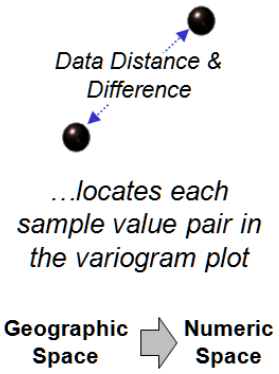


…**Inverse Distance Weighted (IDW)** interpolation uses a fixed $1/D^{Power}$ ***Geometric Equation***

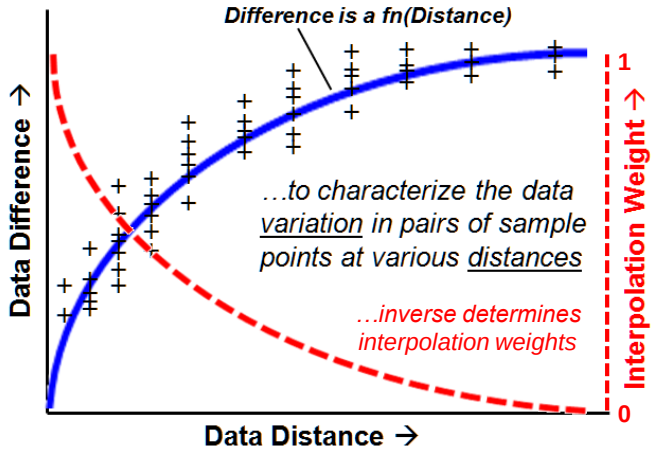
…**Kriging** interpolation uses a ***Derived Equation*** based on regional variable theory (Variogram)

Field Collected Data





Joint Variation



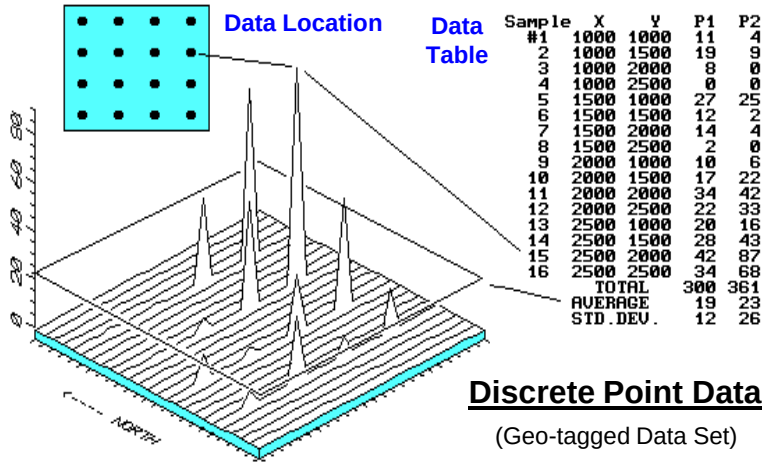
*…instead of a fixed geometric decay function, a **data-driven curve** is derived*

*…and used to determine the **sample weights** used for interpolating each map location*

(Berry)

Spatial Interpolation *(iteratively smoothing the spatial variability)*

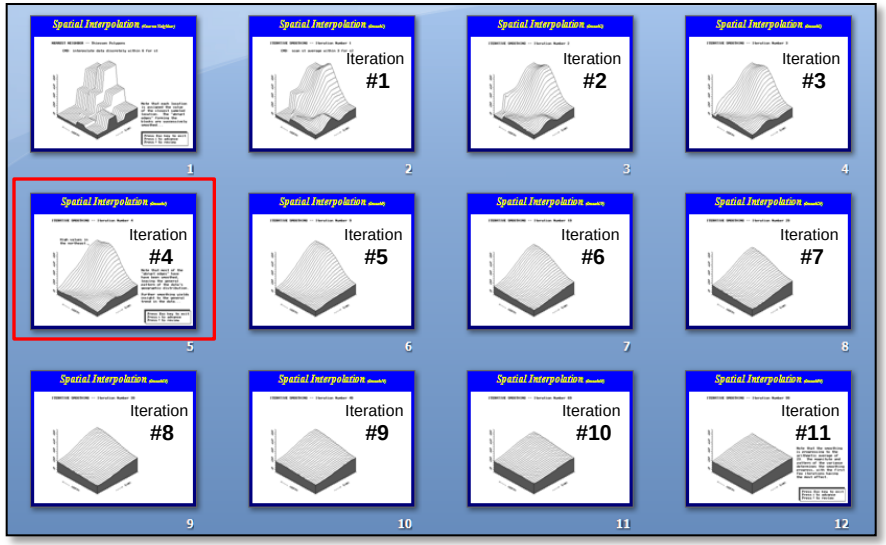
The **iterative smoothing** process is similar to slapping a big chunk of modeler's clay over the "data spikes," then taking a knife and cutting away the excess to leave a continuous surface that encapsulates the peaks and valleys implied in the original data



Sample	X	Y	P1	P2
#1	1000	1000	11	4
2	1000	1500	19	9
3	1000	2000	8	0
4	1000	2500	0	0
5	1500	1000	27	25
6	1500	1500	12	2
7	1500	2000	14	4
8	1500	2500	2	0
9	2000	1000	10	6
10	2000	1500	17	22
11	2000	2000	34	42
12	2000	2500	22	33
13	2500	1000	20	16
14	2500	1500	28	43
15	2500	2000	42	87
16	2500	2500	34	68
	TOTAL		300	361
	AVERAGE		19	23
	STD.DEV.		12	26

Discrete Point Data

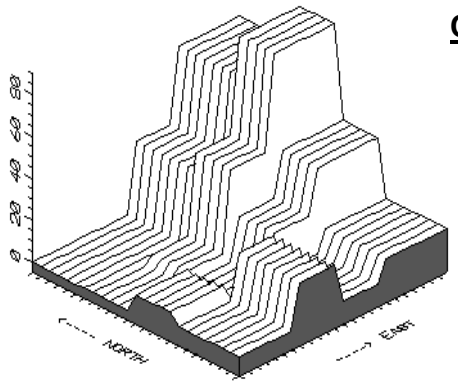
(Geo-tagged Data Set)



(digital slide show SStat2)

NEAREST NEIGHBOR -- Thiessen Polygons

CMD: interpolate data discretely within 6 for s1

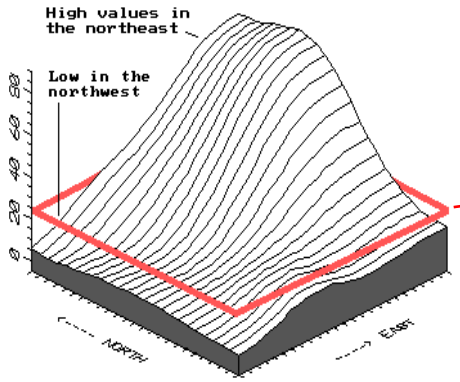


Continuous Surface

Non-sampled locations in the analysis frame are assigned the value of the closest sampled location…

…the "abrupt edges" forming the blocks are iteratively smoothed (**local average**)…

ITERATIVE SMOOTHING -- Iteration Number 4



Valuable insight into the **spatial distribution** of the field samples is gained by comparing the "response surface" with the arithmetic average…

Average value = 23 (± 26)

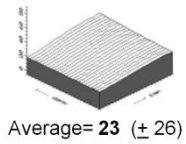
…for each location, its locally implied response is compared to the generalized average

(Berry)

Assessing Interpolation Results *(Residual Analysis)*

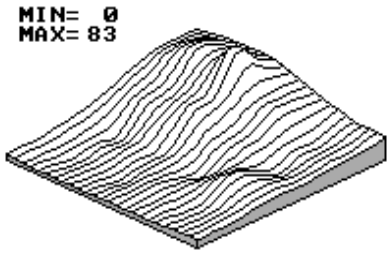
The difference between an actual value (measured) and an interpolated value (estimated) is termed the **Residual**. The residuals can be summarized to assess the performance of different interpolation techniques…

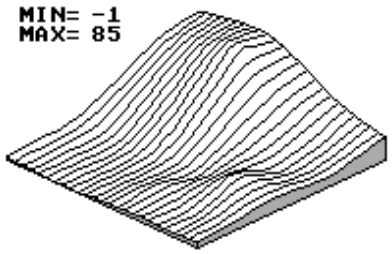
…with the best map surface as the one that has the "**best guesses**" (interpolated estimates)

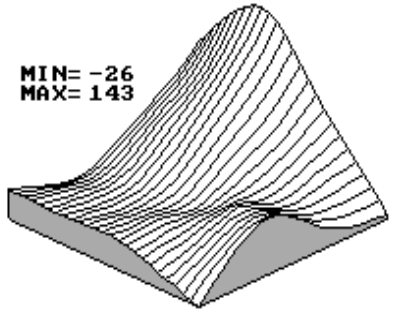


Actual – Estimate = **Residual**

23 – 0 = **23**







Bad Guess

Sample	Col	Row	Actual	Average	Inverse	Kriging	MinCurve
#17*	1	1	0	23 (23)	8 (8)	2 (2)	6 (6)
18*	18	2	48	23 (-25)	42 (-6)	46 (-2)	28 (-20)
19*	23	2	64	23 (-41)	52 (-12)	65 (1)	65 (1)
20*	19	4	65	23 (-42)	54 (-11)	56 (-9)	48 (-17)
21	15	6	34	23 (-11)	33 (-1)	30 (-4)	31 (-3)
22	5	7	0	23 (23)	2 (2)	-1 (-1)	1 (1)
23	9	8	6	23 (17)	7 (1)	1 (-5)	1 (-5)
24	19	11	79	23 (-56)	67 (-12)	70 (-9)	69 (-10)
25	23	13	64	23 (-41)	52 (-12)	68 (4)	90 (26)
26*	4	16	8	23 (15)	8 (0)	7 (-1)	6 (-2)
27	16	17	19	23 (4)	22 (3)	19 (0)	17 (-2)
28*	2	20	6	23 (17)	8 (2)	3 (-3)	-6 (-12)
29	13	22	12	23 (11)	15 (3)	14 (2)	19 (7)
30	22	22	17	23 (6)	19 (2)	20 (3)	7 (-10)
31*	2	24	9	23 (14)	8 (-1)	6 (-3)	-16 (-25)
32	19	24	14	23 (9)	19 (5)	11 (-3)	-7 (-21)
Test Set Average= 28							
Average Estimate=				23	26	26	22
Sum of the Residuals=				(-77)	(-29)	(-28)	(-86)
Average Unsigned Residual=				(22.2)	(5.1)	(3.3)	(10.5)
Normalized Residual Index=				(.80)	(.18)	(.12)	(.38)

Best Surface

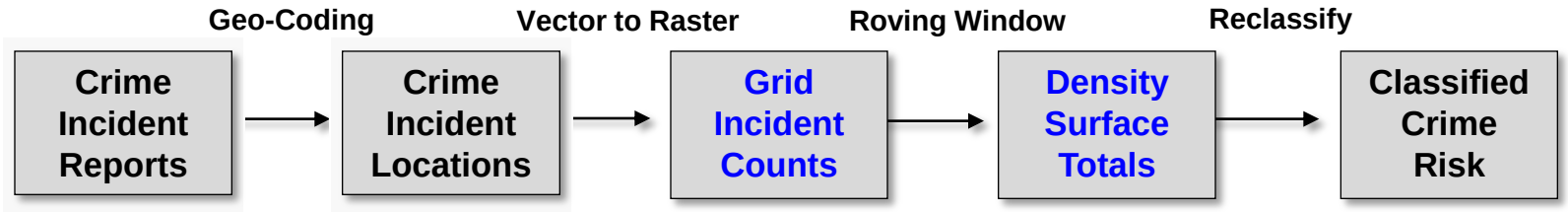
(Berry)

Creating a Crime Risk Density Surface *(Density Analysis)*

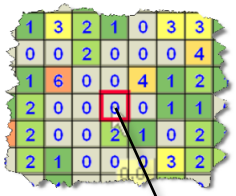
Spatial Statistics:

Basic Descriptive Statistics
Basic Classification
Map Comparison
Unique Map Statistics
Surface Modeling
Advanced Classification
Predictive Statistics

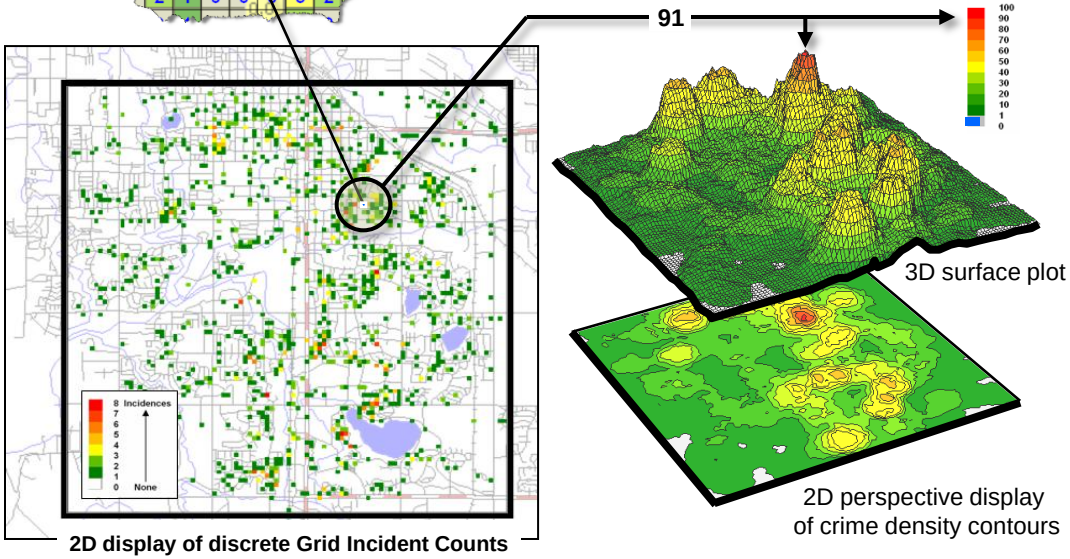
Density Analysis "counts or sums" data values within a specified distance from each map location (roving window) to generate a continuous surface identifying the relative spatial concentration of data within a project area, such as the number of customers or bird sightings within a half mile.



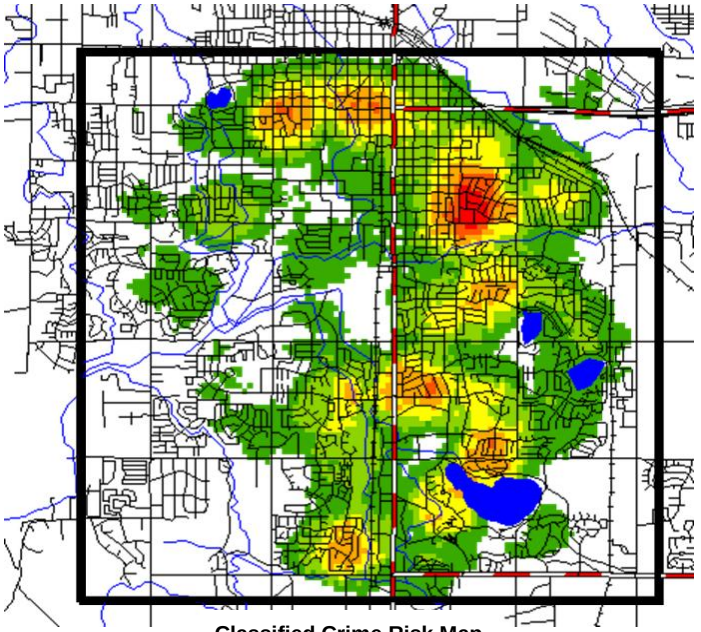
Geo-coding identifies geographic coordinates from street addresses



Grid Incident Counts
the number of incidences (points) within in each grid cell

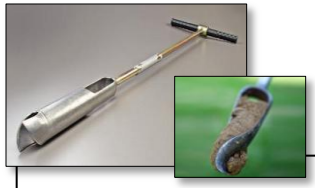


Calculates the total number of reported crimes within a roving window– ***Density Surface Totals***



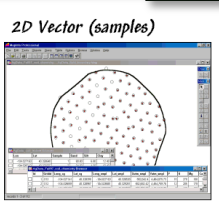
(Berry)

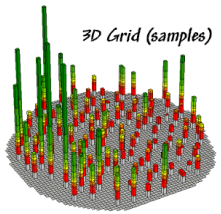
Visualizing Spatial Relationships

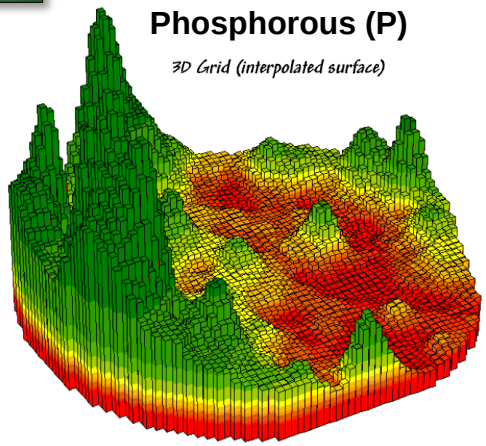


Interpolated Spatial Distribution

of soil nutrient concentrations







Multivariate Coincidence

What spatial relationships do you see?

…do relatively high levels of P often occur with high levels of K and N?

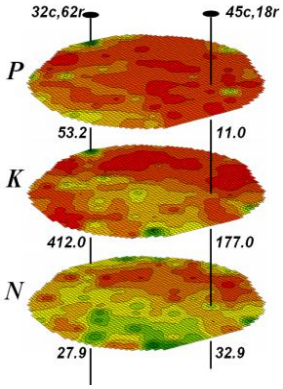
…how often?

…where?

Humans can only "see" broad

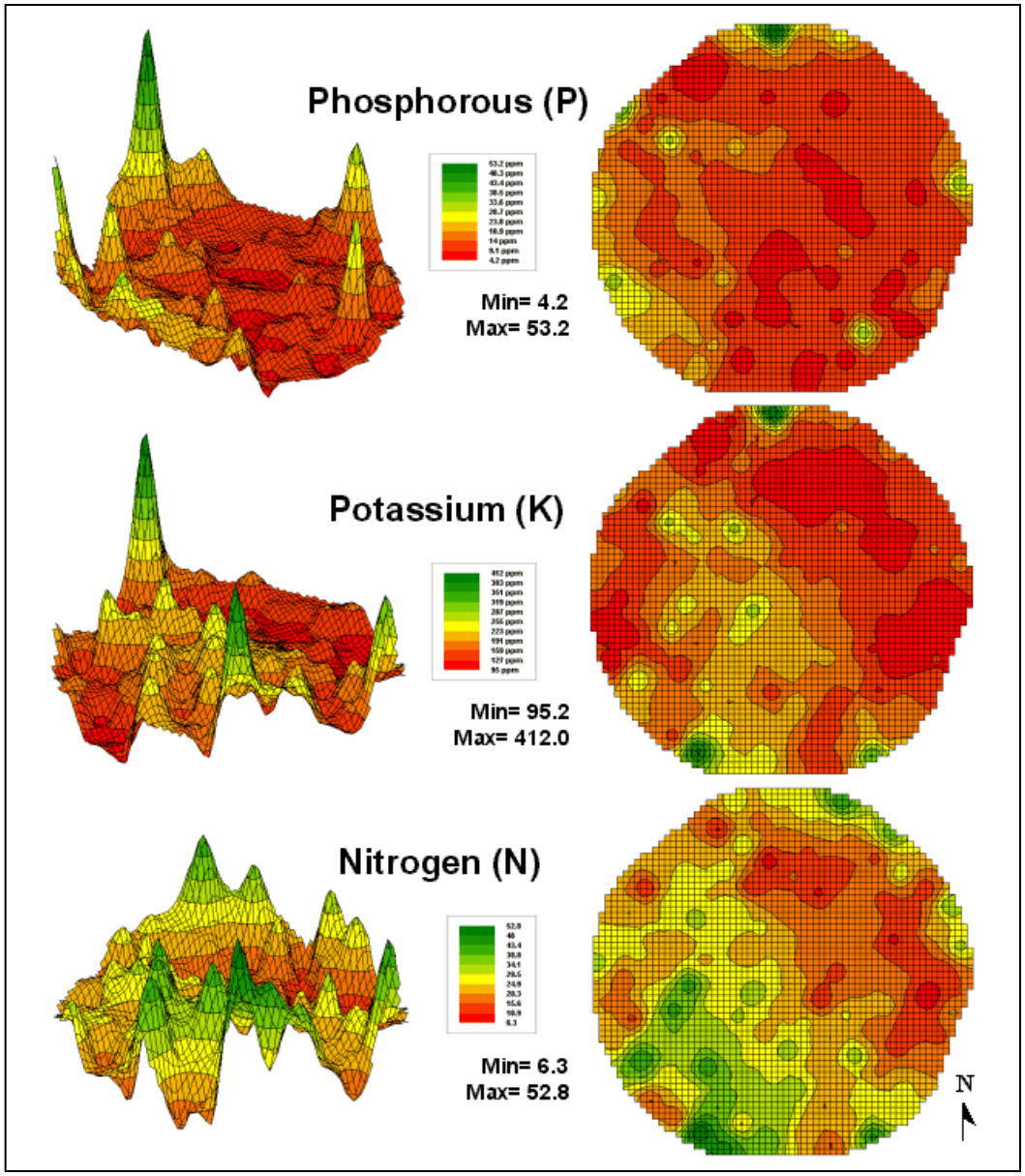
Generalized Patterns

in a single map variable



(Map Stack)

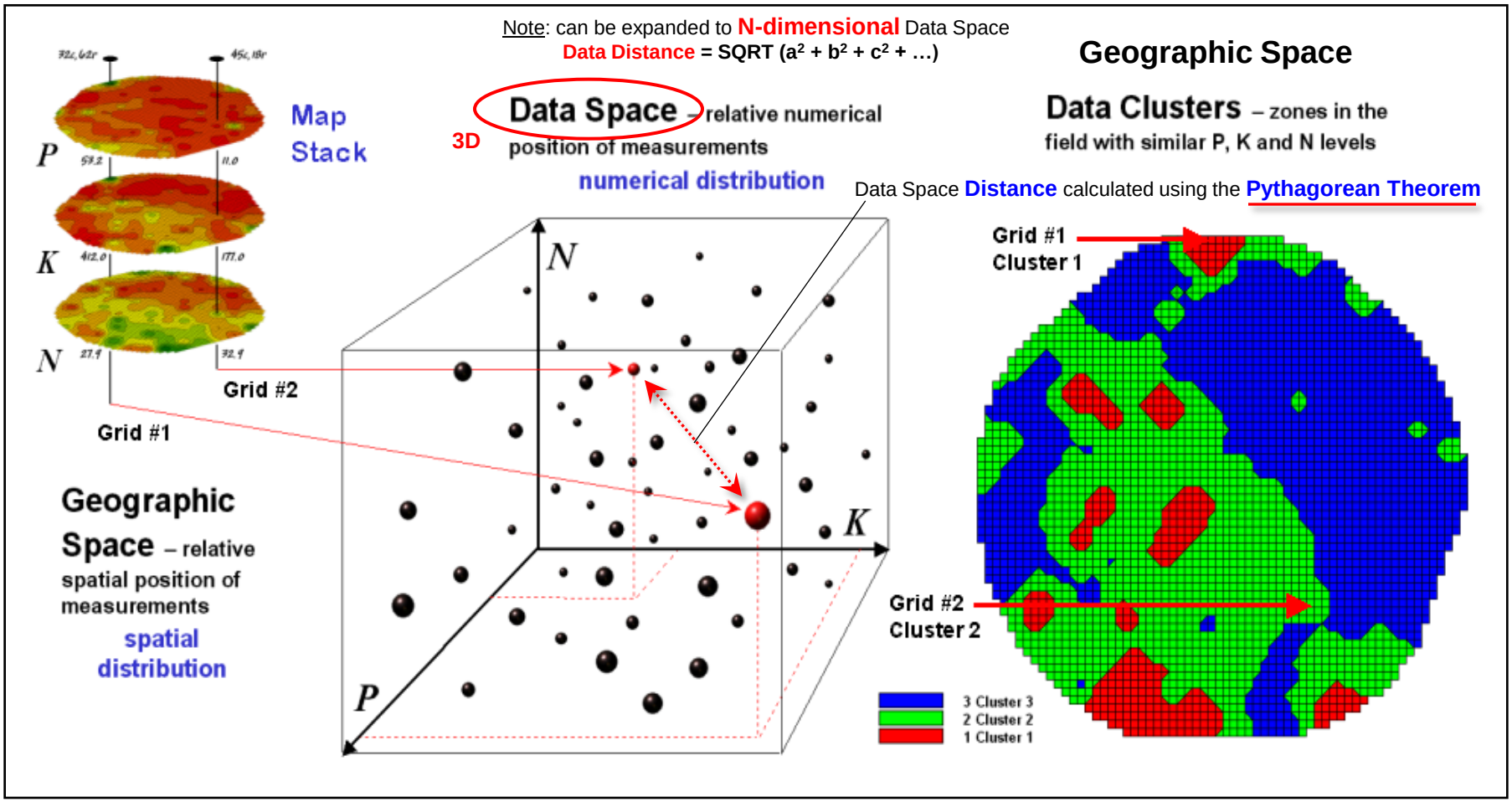
Continuous Grid-Map Surfaces *(Data Layers)*



(Berry)

Clustering Maps for Data Zones

…but *computers* can "see" detailed *numeric patterns* in *multiple* map variables using ***Data Space***



…groups of relatively close "floating balls" in data space identify locations in the field with similar data patterns– ***Data Zones*** *(Data Clusters)*

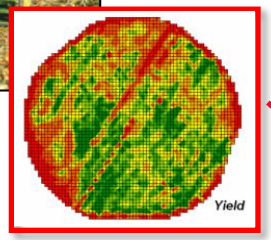
…the IsoData algorithm ***minimizes Intra-Cluster distances*** *(within a cluster— similar) while at the same time* ***maximizing Inter-Cluster distances*** *(between clusters— different)*

(Berry)

The Precision Ag Process

On-the-Fly
Yield Map

Steps 1–3)



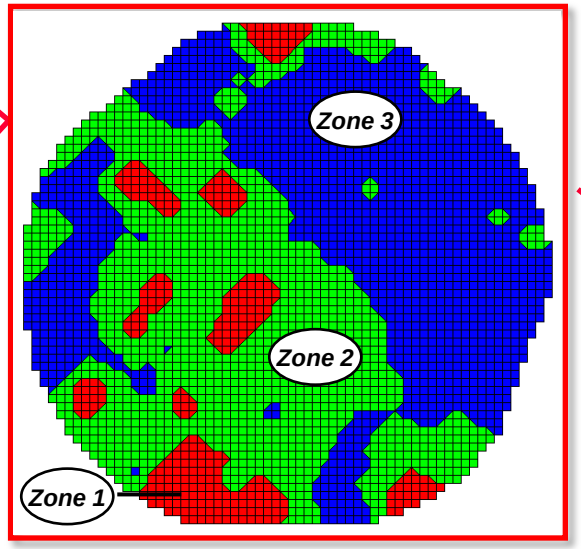
As a combine moves through a field it…

*1) uses **GPS to check its location** every second then*

*2) records the **yield monitor value** at that location to*

*3) create a continuous **Yield Map surface** identifying the variation in crop yield every few feet throughout the field (**dependent map variable**).*

Step 5) Prescription Map → ***"As-applied" maps***



Intelligent Implements

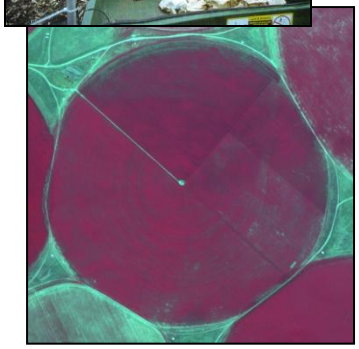
Step 6)

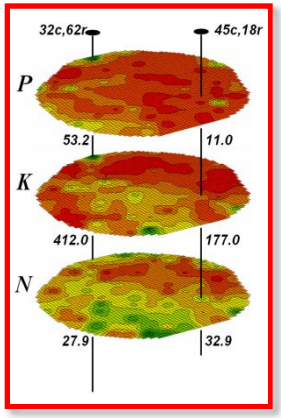
Variable Rate Application

*4) …**soil samples** are interpolated for continuous **Nutrient Map surfaces**.*

Step 4)

*Derived **Soil Nutrient Maps***





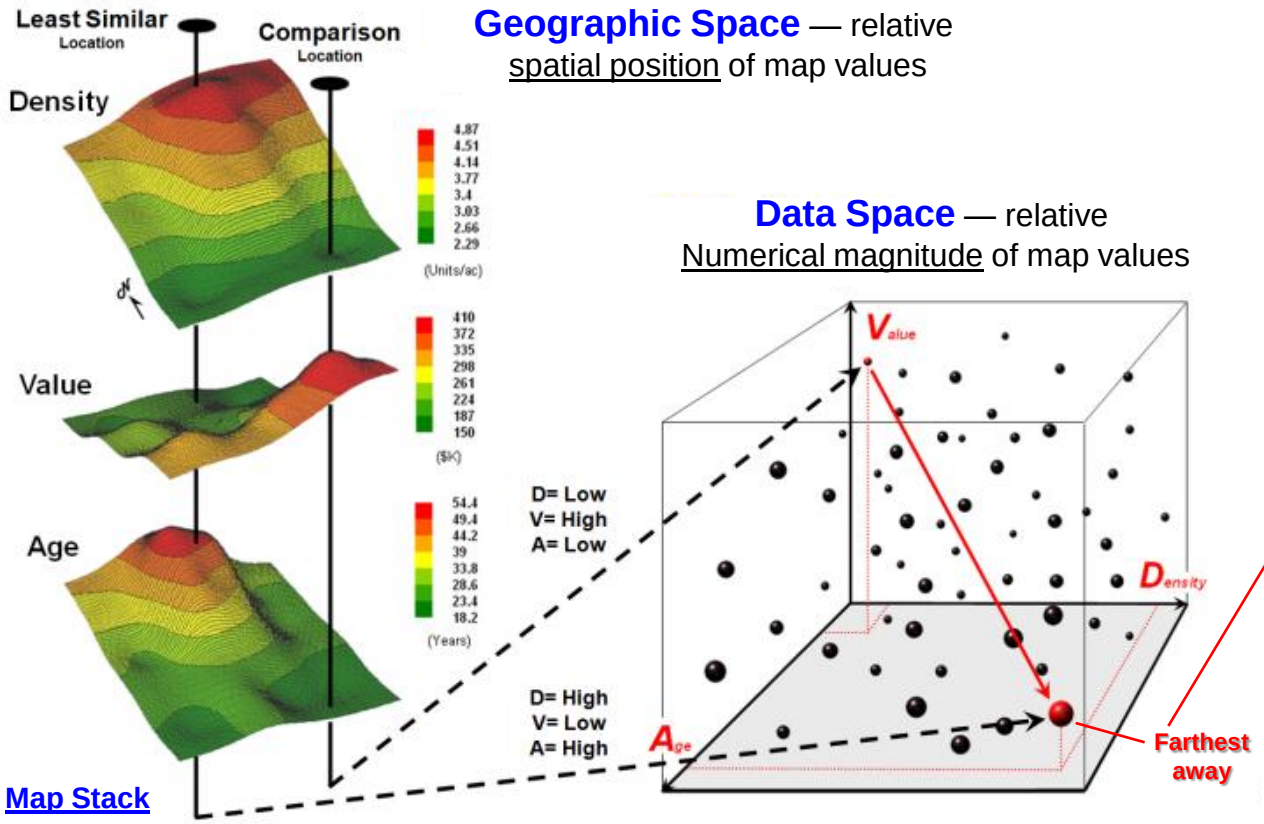
*5) The **yield map is analyzed in combination with soil nutrient maps**, terrain and other mapped factors (**independent map variables**) to derive a **Prescription Map**…*

*6) …that is used to **adjust fertilization levels** applied every few feet in the field (**If <condition> then <action>**).*

…more generally termed the **Spatial Data Mining Process** (e.g., Geo-Business application)

(Berry)

Map Similarity *(identifying similar numeric patterns)*



Geographic Space — relative spatial position of map values

Data Space — relative Numerical magnitude of map values

Locations identical to the Comparison Point are set to 100% similar

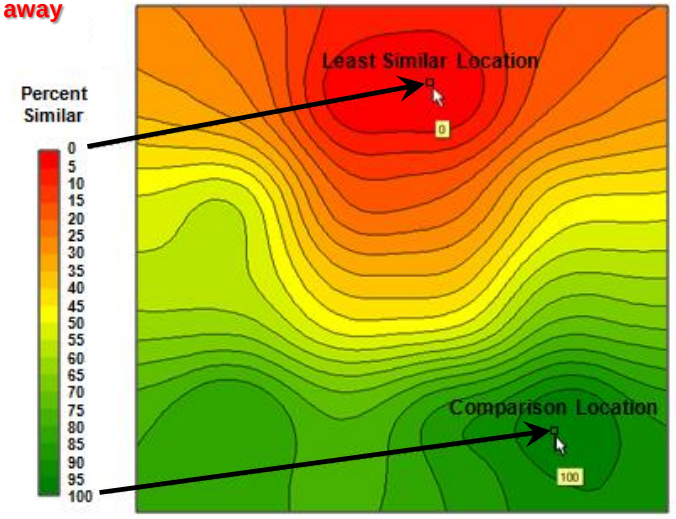
*(**Identical** numerical pattern)*

The farthest away point in data space is set to 0

*(**Least Similar** numerical pattern)*

…all other Data Distances are scaled in terms of their relative similarity to the comparison point

*(**0 to 100% similar**)*



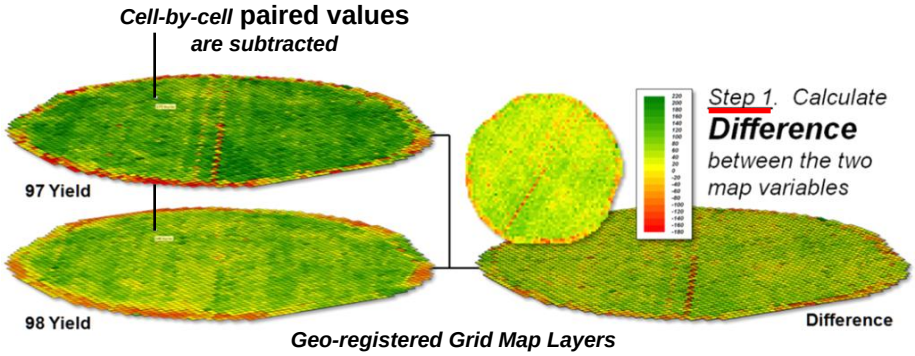
Each "floating ball" in the **Data Space** scatter plot schematically represents a location in the field (**Geographic Space**).

The position of a ball in the plot identifies the relative phosphorous (P), potassium (K) and nitrogen (N) levels at that location.

Basic Descriptive Statistics
Basic Classification
Map Comparison
Unique Map Statistics
Surface Modeling
Advanced Classification
Predictive Statistics

(Berry)

Map Comparison *(spatially evaluating the T-test)*

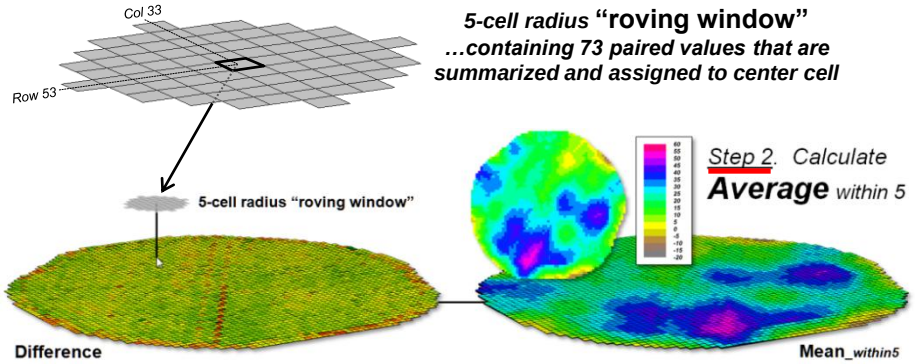


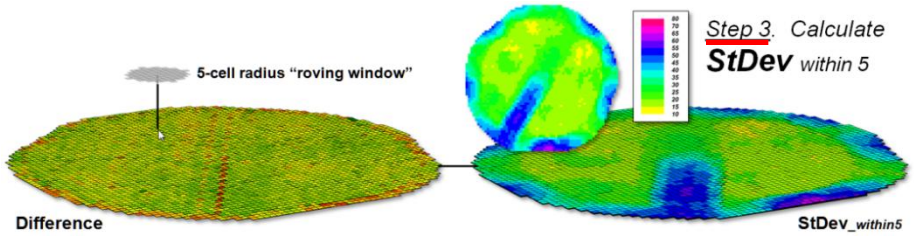
Spatially Evaluating the "T-Test"

*The **T-statistic** equation is evaluated by first calculating a map of the **Difference** (Step 1) and then calculating maps of the **Mean** (Step 2) and **Standard Deviation** (Step 3) of the Difference within a "roving window." The **T-statistic** is calculated using the derived Mean and StDev maps of the localized difference using the equation (step 4) —*

spatially localized solution

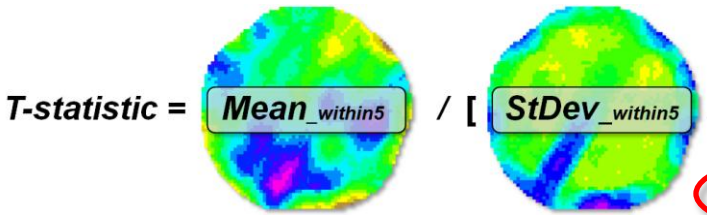
$$\text{T-statistic} = \frac{\text{Mean}_{\text{difference}}}{\text{StDev}_{\text{difference}} / \text{Sqrt}\,(73)}$$



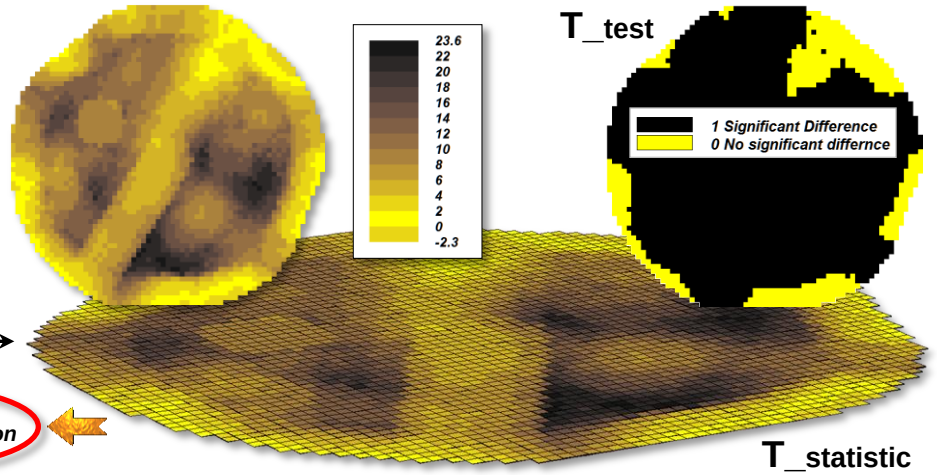


Step 4. Calculate the ***"Localized" T-statistic***
(using a 5-cell roving window) for each grid cell location

*…the result is map of the **T-statistic** indicating how different the two map variables are throughout geographic space and a **T-test** map indicating where they are significantly different.*



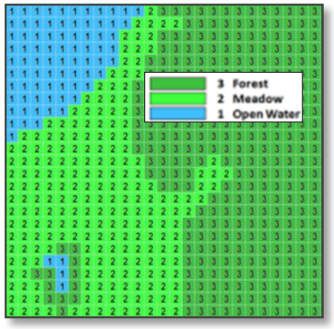
Evaluate the Map Analysis Equation



(Berry)

Spatial Statistics Operations *(Numerical Context)*

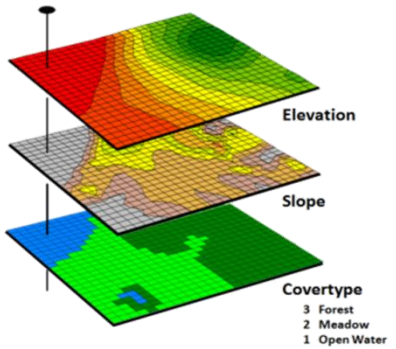
GIS as "Technical Tool" *(Where is What)* vs. "**Analytical Tool**" *(Why, So What and What if)*



Grid Layer

Spatial Statistics

Map Stack



Spatial Statistics seeks to map the spatial variation in a data set instead of focusing on a single typical response (central tendency) ignoring the data's spatial distribution/pattern, and thereby provides a mathematical/statistical framework for *analyzing* and *modeling* the

Numerical Spatial Relationships

within and among grid map layers

…discussion focused on these groups of spatial statistics — see reading references for more information on all of the operations

Statistical Perspective:

Map Analysis Toolbox

Basic Descriptive Statistics *(Min, Max, Median, Mean, StDev, etc.)*

Basic Classification *(Reclassify, Contouring, Normalization)*

Map Comparison *(Joint Coincidence, Statistical Tests)*

✓ ***Unique Map Statistics*** *(Roving Window and Regional Summaries)*

✓ ***Surface Modeling*** *(Density Analysis, Spatial Interpolation)*

Advanced Classification *(Map Similarity, Maximum Likelihood, Clustering)*

Predictive Statistics *(Map Correlation/Regression, Data Mining Engines)*

(Berry)