
Spatial Analysis for Water Resources Modeling and Management :

Methods for Analysis, Interpretation and Visualization of Spatial Data

Ramesh Teegavarapu, Ph.D., P.E.
Associate Professor,
Director, Hydrosystems Research Laboratory (HRL)
Department of Civil, Environmental and Geomatics Engineering,
Florida Atlantic University, Boca Raton, Florida, 33431, USA

Permission to use.

- The material in the presentation is obtained from several copyright protected sources (including journal publications, books and published articles, technical presentations by author and his co-authors). Permission to use the material in this presentation elsewhere needs to be obtained from the author(s) of this presentation as well as publishing agencies which own the copyright permissions for the figures and illustrations.
- Material in the presentation can only be for Academic Use only.
- Journal articles for personal use can be obtained from author: rteegava@fau.edu
- Some of figures are not yet published in any article by the author.
- Any additional information please contact the author :
rteegava@fau.edu

HRL

Water . Environment . Climate .

*Committed to Understanding, Modeling
and Managing Terrestrial Hydro-
Environmental Systems*



HYDROSYSTEMS RESEARCH LABORATORY

Home Vision Research **Projects** Published Works Presentations Work Space Innovation Impact Personnel Support
Opportunities Contact News Hydroanalytics

<http://hrl.fau.edu>

Vision

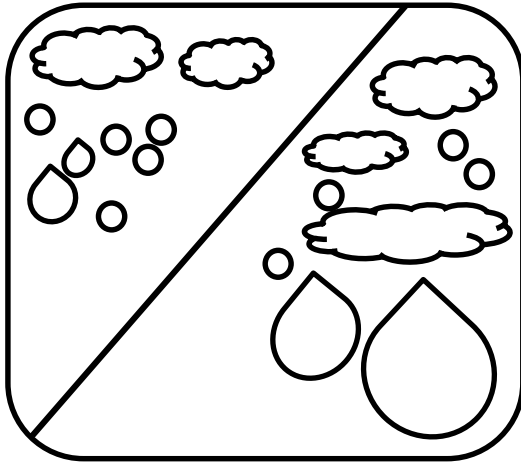
HRL promotes development and application of seminal and original research methodologies and tools for understanding, analyzing, modeling and managing our hydro and environmental-systems specific terrestrial environment.



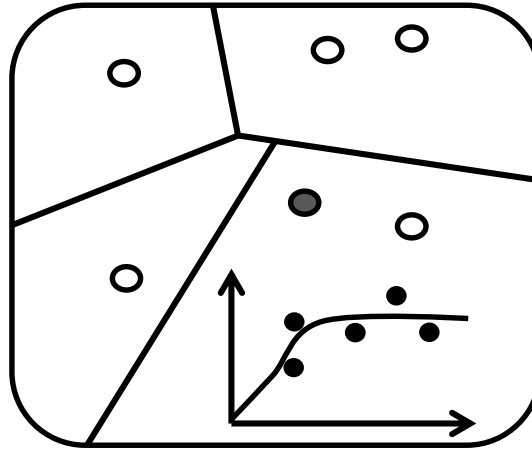
Research Studies



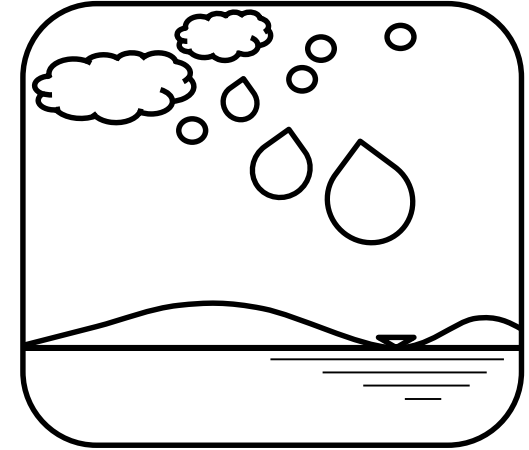
Climate Change . Climate Variability . Water Resources Systems. Watershed Modeling . Hydrologic Modeling . Precipitation Processes . Stormwater Modeling and Management . Monitoring Networks . Hydrometeorology . Statistical Hydrology . Decision Support Systems . Hydroinformatics .



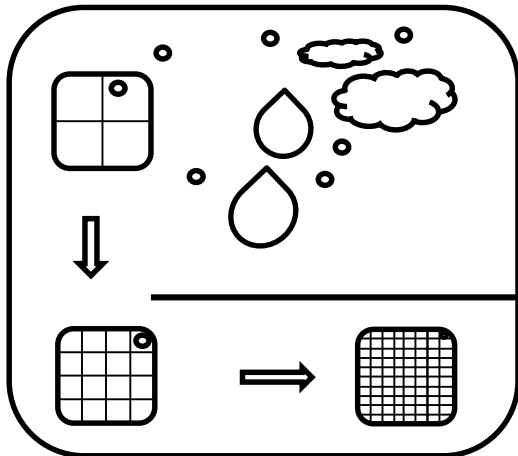
**Extreme Precipitation and
Climate Change**



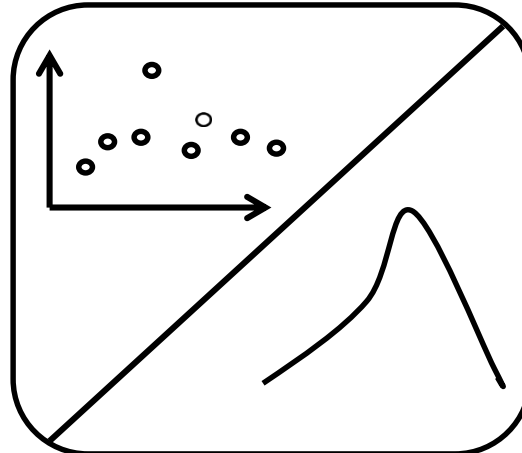
**Spatial Interpolation
Surface Interpolation, Missing Data**



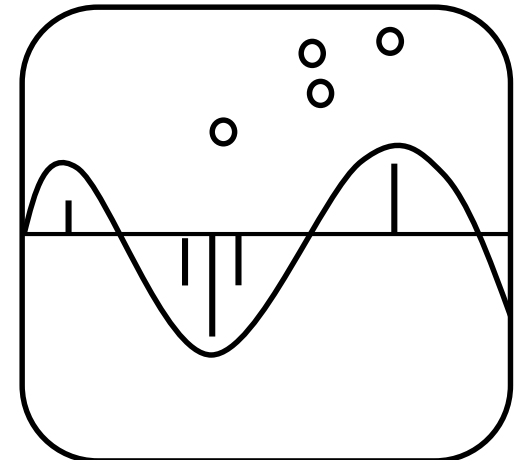
**Extreme Precipitation
and Floods**



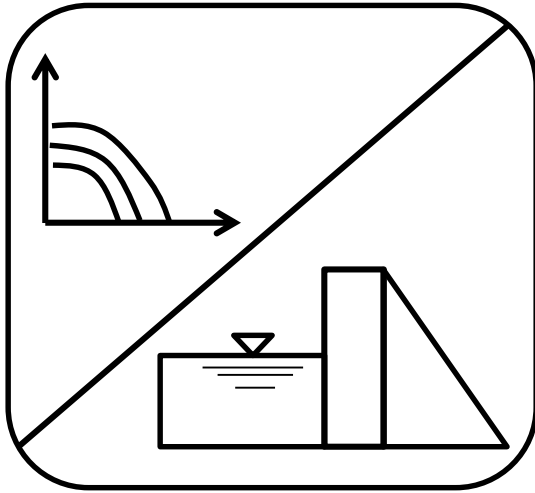
**Climate
Downscaling Methods**



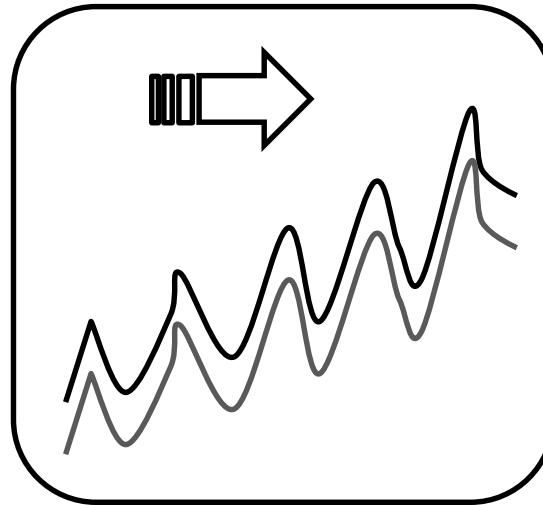
**Statistical Evaluation of Extremes
Climate Change Trends**



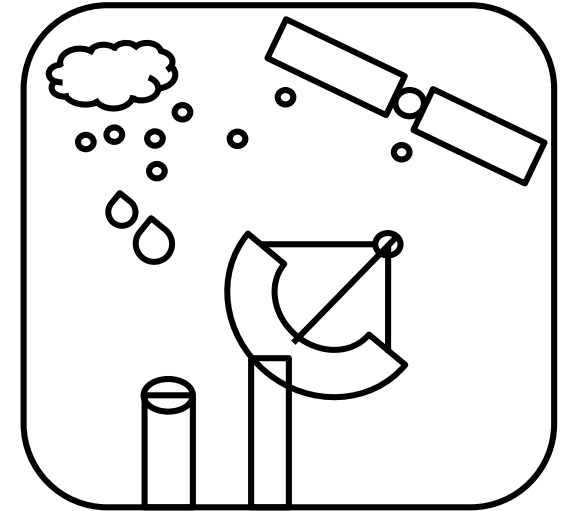
**Climate Variability
Teleconnections**



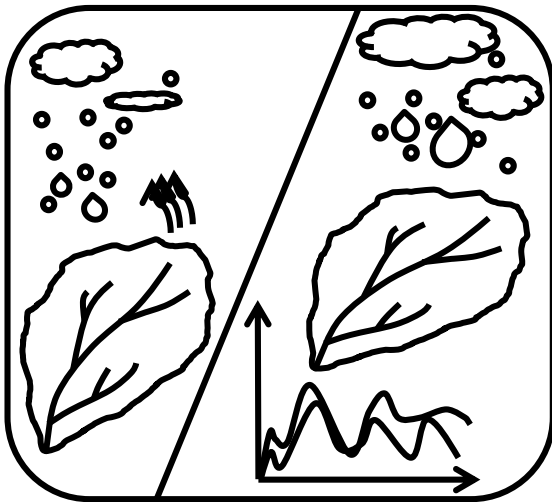
Hydrologic Design and Water Resources Management under Climate Change



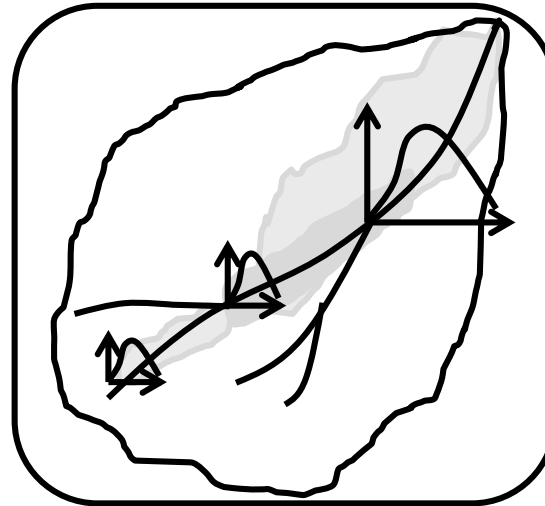
Climate Change: Future Trends



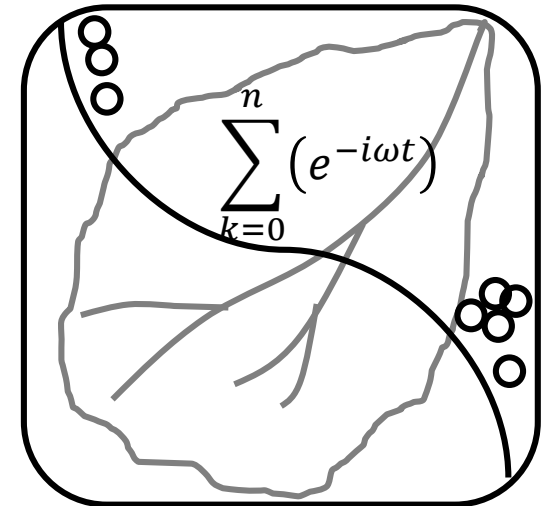
Satellite, Radar, Rain gage Estimation & Algorithms



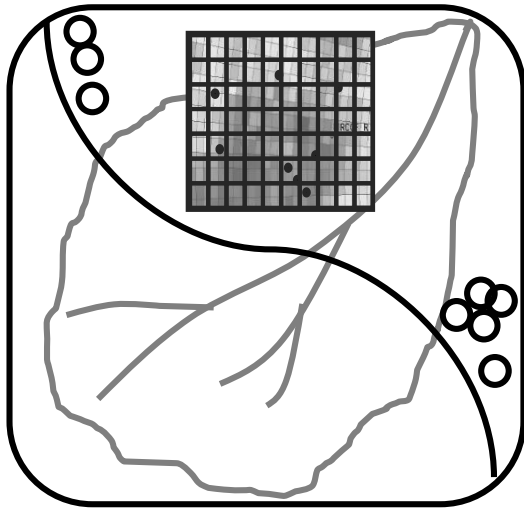
Watershed Modeling/Water Quality Fate and Transport Modeling



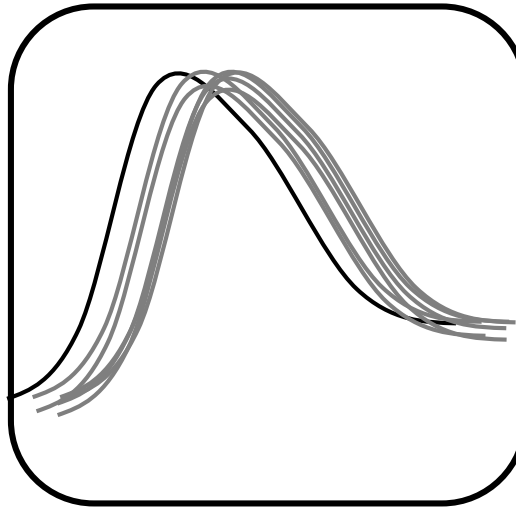
Catastrophic Floods Animation/Visualization



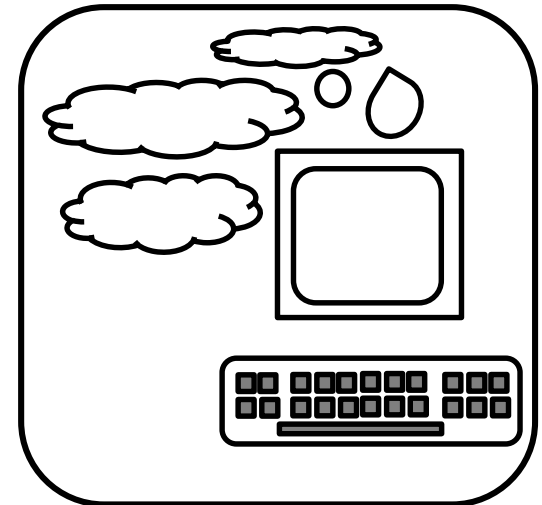
Water Resources/Environmental Optimization Models



**Spatial Analysis & GIS
(Hydrological & Environmental)**



**Uncertainty Analysis,
Ensemble Forecasting,
Data Assimilation**



**Environmental
Intelligent/Knowledge-
based
Decision Support Systems**

Supporting Agencies

NSF (National Science Foundation)

FEMA (Federal Emergency Management Agency)

FAU (Florida Atlantic University)

FSUS(Florida State University System)

Dewberry (Dewberry.com)

Jupiter's Last Call



FEMA



FAU
FLORIDA
ATLANTIC
UNIVERSITY



Hinkley Center for Solid and Hazardous Waste Management

USGS (United States Geological Survey)

SFWMD (South Florida Water Management District)

SWFWMD (Southwest Florida Water Management District)

FDOT (Florida Department of Transportation)

FDEP (Florida Department of Environmental Protection)

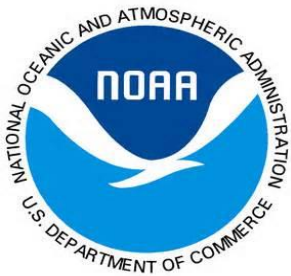
Research work at HRL is supported by several funding/sponsoring agencies .



Supporting Agencies

NOAA (PRIDE program of National Oceanic and Atmospheric Administration)

Manitoba Hydro

NSERC (Natural Sciences and Engineering Research Council)




 Natural Sciences and Engineering
 Research Council of Canada
 
 Conseil de recherches en sciences
 naturelles et en génie du Canada



Total Research Funding : Over 3 million dollars (Secured by Dr. T. At HRL as Principal Investigator or Co-Principal Investigator working on research teams at HRL since January 2007).



USGS (United States Geological Survey)

Kentucky Division of Water (KDOW)

Environmental Protection Agency (EPA)

Description of the Course

- This course presents a comprehensive review of spatial analysis methods and tools that are critical for analysis, modeling and management of hydrologic, environmental and water resource systems. The course emphasizes on three issues: (1) analysis, (2) methodologies and (3) techniques.
- The course provides an exhaustive background and in-depth review of spatial point patterns, statistics, and geographical information systems and discusses methods to evaluate spatial information, clusters, interpolation methods, generation of spatially varying inputs to distributed hydrologic models, and approaches for hydrometeorological monitoring network design.
- Spatial analysis methods and their applications for watershed and water resources management will be discussed and elaborated in this course.

Course Topics

- Introduction to Spatial Analysis
- Introductory overview of spatial analysis. Observations in space and development of sampling schemes. Modifiable areal unit problem (MAUP), spatial autocorrelation, scales of measurement, extents. Local and global models: regression, moving window regression, geographically weighted regression.

DATA

- Types of Data
 - Temporal
 - Spatial
- What are the differences between spatial and temporal data ?

Exploratory Data Analysis (EDA)

- EDA is an approach or philosophy that employs a variety of techniques
 - Maximize insight into a data set;
 - Uncover underlying structure;
 - Extract important variables;
 - Detect outliers and anomalies;
 - Test underlying assumptions;
 - Develop parsimonious models; and
 - Determine optimal factor settings.

According to the NIST handbook

Steps	Description
1. Problem definition	<p>Define</p> <ul style="list-style-type: none"> ● Objectives ● Deliverables ● Roles and responsibilities ● Current situation ● Timeline ● Costs and benefits
2. Data preparation	<p>Prepare and become familiar with the data:</p> <ul style="list-style-type: none"> ● Pull together data table ● Categorize the data ● Clean the data ● Remove unnecessary data ● Transform the data ● Partition the data
3. Implementation of the analysis	<p>Three major tasks are</p> <ul style="list-style-type: none"> ● Summarizing the data ● Finding hidden relationships ● Making prediction <p>Select appropriate methods and design multiple experiments to optimize the results. Methods include</p> <ul style="list-style-type: none"> ● Summary tables ● Graphs ● Descriptive statistics ● Inferential statistics ● Correlation statistics ● Searching ● Grouping ● Mathematical models
4. Deployment	<ul style="list-style-type: none"> ● Plan and execute deployment based on the definition in step 1 ● Measure and monitor performance ● Review the project

Missing Data

- Missing data exists in almost all datasets
- Missing data needs to be identified and tagged.
- Identification requires assessment of missing data mechanisms

Handling Missing Data

- Assume them to be not available or not part of the record and carry out the analysis with rest of the data.
- Use temporal or spatial interpolation to estimate missing values.
- Temporal interpolation requires a very strong autocorrelation (a strong dependence of value in time “ t ” on the value in time “ $t-1$ ”)
- If the temporal autocorrelation is very weak, temporal interpolation cannot be used.
- Temporal interpolation may be based on
 - Average of two values (from time “ $t-1$ ” and “ $t+1$ ”, with observation missing in time “ t ”)
 - A variant would be to use linear or non-linear interpolation

All about Data

- **Babbage, Charles (1792-1871)**
Errors using inadequate data are much less than those using no data at all.
- **Doyle, Sir Arthur Conan (1859-1930)**
It is a capital mistake to theorize before one has data. *Scandal in Bohemia*.
- Although we often hear that data speak for themselves, their voices can be soft and sly.
— Frederick Mosteller
Beginning Statistics with Data Analysis (1983),
234

Tools

- [S-PLUS](#) - from Insightful Corp. (formerly MathSoft, StatSci)
- [SPSS](#)
- [Stata](#)
- [STATISTICA](#) - from StatSoft
- [STATGRAPHICS Centurion XVI](#)
- [Statistical Solutions \(nQuery, BMDP, NCSS\)](#)
- [Statistix](#)
- [Systat](#)
- [UNISTAT Statistical Package](#)
- [WINKS](#) - (formerly KWIKSTAT) from TexaSoft
- [XploRe](#)
- [Excel](#) – Data Analysis Pack

Tools

- [Maple](#) - from Waterloo Maple
- [Matlab](#) - from Mathworks
- [Mathematica](#) - from Wolfram Research
- [Mathcad](#) - from Mathsoft
- [mathStatica](#) - mathematical statistics with Mathematica

Tools

- [BioConductor](#) - software for bioinformatics
- [ESS](#) - Emacs Speaks Statistics
- [GGobi Data Visualization System](#)
- [The Omega Project for Statistical Computing](#)
- [The R Project for Statistical Computing](#)
- [ROOT](#) - an object-oriented data analysis framework
- [Statistics Online Computational Resource \(SOCR\)](#)
- [StatLib](#) - a system for distributing statistical software, data sets, and information
- [ViSta](#) - The Visual Statistics System
- [Weka Machine Learning Project](#)



MATLAB

- MATLAB is a proprietary software (commercial software) from MATHWORKS.com
- A complete computing package
- MATLAB : **MAT**rix **LAB**oratory
- Uses a simple programming language – similar to C syntax, easy to learn
- Uses Interpreter – executes commands in order
- Provides ability to communicate with Web-based services
- Provides ability to develop graphical user interfaces (GUIs) for front-end applications •
- Script files and functions can be written, saved and executed.
- Comprehensive Interfaces can be developed using MATLAB for data handling and generation of visual results and input manipulation.
- Thousands of user provided scripts are provided (un tested for reliability)

Tool Boxes

- MATLAB uses a number of “Tool Boxes” to carry out different calculations.
- Not all functions are available in one particular “Tool Box”
- Examples of “Tool Boxes” :
 - Statistical tool box
 - Curve fitting tool box
 - Econometrics tool box
 - Neural Networks tool box
 - Signal Processing tool box
 - Fuzzy Set tool box
 - Database tool box
 - Mapping tool box

MATLAB Data Input

- Data needs to be in matrix format
- Matrices that are sparse are not accepted
- Data can be read from Excel, text or .CSV files
- Complex data format and strings can be read
- No limitation on the size of the matrix.
- Three dimensional matrices can also be read

Wessa

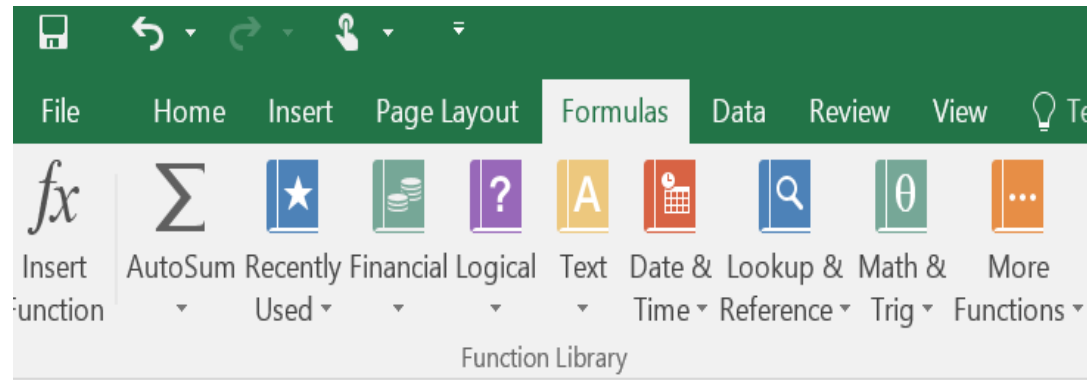
- Wessa (<http://www.wessa.net/>) is a free online statistics applications for analyzing statistical data.
- The data can be provided through a web-based interface and the analysis is reported online.
- The Wessa.net basically takes your data and runs a “R” code in the background and provides the results.
- Graphics are also generated and can be exported.
- Wessa.net also provides datasets for analysis.
- If you use Wessa.net, the web site asks that you credit the software.

Wessa

- Wessa (<http://www.wessa.net/>) is an application software for analyzing statistical data.
 - Visualization
 - Statistical inference
- The data can be provided through a web-based interface and the analysis is reported online.
- The Wessa.net basically takes your data and runs a “R” code in the background and provides the results.
- Graphics are also generated and can be exported.
- Wessa.net also provides datasets for analysis.
- If you use Wessa.net, the web site asks that you credit the software.

Excel Functions

- Compatibility functions
- Cube functions
- Database functions
- Date and time functions
- Engineering functions
- Financial functions
- Information functions
- Logical functions



Excel Functions

- Lookup and reference functions
- Math and trigonometry functions
- Statistical functions
- Text functions
- User defined functions that are installed with add-ins
- Web functions

Add-in Functions/Tools

- Goal Seek
 - A single variable equation solver.
 - Very handy for many engineering applications.
 - Solves the best variable value considering an equation format.
 - Multiple solutions are not possible.

Example:

Solve for r , when A is given

$$A = \pi r^2 + 9r^3 + r$$

Add-on Functions/Tools

- Solver
 - Can be used to solve linear and nonlinear optimization problems.
 - Uses both derivative-based and evolutionary programming approaches for obtaining optimal solutions
 - Can solve mixed integer problems along with binary variables
 - An optimization problem will involve:
 - Objective function
 - Constraints (equality, inequality, hard and soft)

EDA and Other Analysis Techniques

- Three popular data analysis approaches are:
 1. Classical
 - 2. Exploratory (EDA)**
 3. Bayesian

Exploratory Data Analysis

- **Exploratory Data Analysis (EDA)** is an approach/philosophy for data analysis that employs a variety of techniques (**mostly graphical**) to maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- determine optimal factor settings.

Sequence of Analysis in Different Methods

- **Classical analysis**

- Problem \Rightarrow Data \Rightarrow Model \Rightarrow Analysis \Rightarrow Conclusions

- **EDA**

- Problem \Rightarrow Data \Rightarrow Analysis \Rightarrow Model \Rightarrow Conclusions

- **Bayesian**

- Problem \Rightarrow Data \Rightarrow Model \Rightarrow Prior Distribution \Rightarrow Analysis \Rightarrow Conclusions

Models: Inductive & Deductive

Inductive Models

- Inductive models are basically data-driven models.
- No understanding of the physics of the process is attempted
- Data (observed or surrogate data associated with a variable/variables) are used to build models using function approximation tools.

Deductive Models

- Inductive Models – Induction
- Deductive Models – Deduction
- Deductive Modeling
 - Models are built based on complete understanding of the physical processes that influence the system
 - Mathematical relationships (dimensionally homogeneous) developed.
 - The relationships are evaluated using available data

Inductive Vs Deductive Models

- Inductive – computationally inexpensive, data-driven, model dependent on the methods used. Empirical, case-study specific and data-sensitive and intensive.
- Deductive – computationally expensive, models derived based on knowledge of the system, universal relationships that are not case-study specific.

Empirical Models : Issues

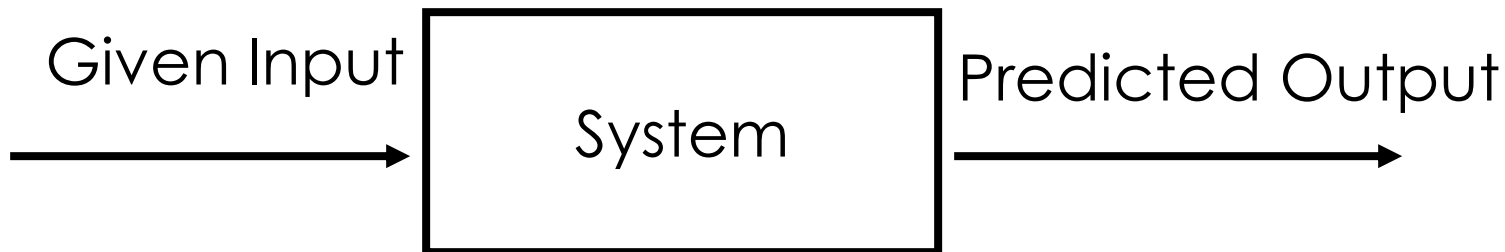
- Over-fitting
- Under-fitting
- Use of wrong functional form for fitting the data
- Spurious associations (correlations)
- Selection of the variables (predictors) in a predictor-predictand relationship.
- Generalization
- Interpolation and Extrapolation

Occam's Razor

- Principle of Parsimony
Choose the simplest model
- The principle states that one should not make more assumptions than the minimum needed.
- In any given model, Occam's razor helps us to "shave off" those concepts, variables or constructs that are not really needed to explain the phenomenon

Descriptive Models

- Scientists and engineers develop and use descriptive models for the purpose of describing a physical system or sub-system and for the purpose of predicting the behavior of the system in response to a given stimulus or loading:



Inductive Modeling

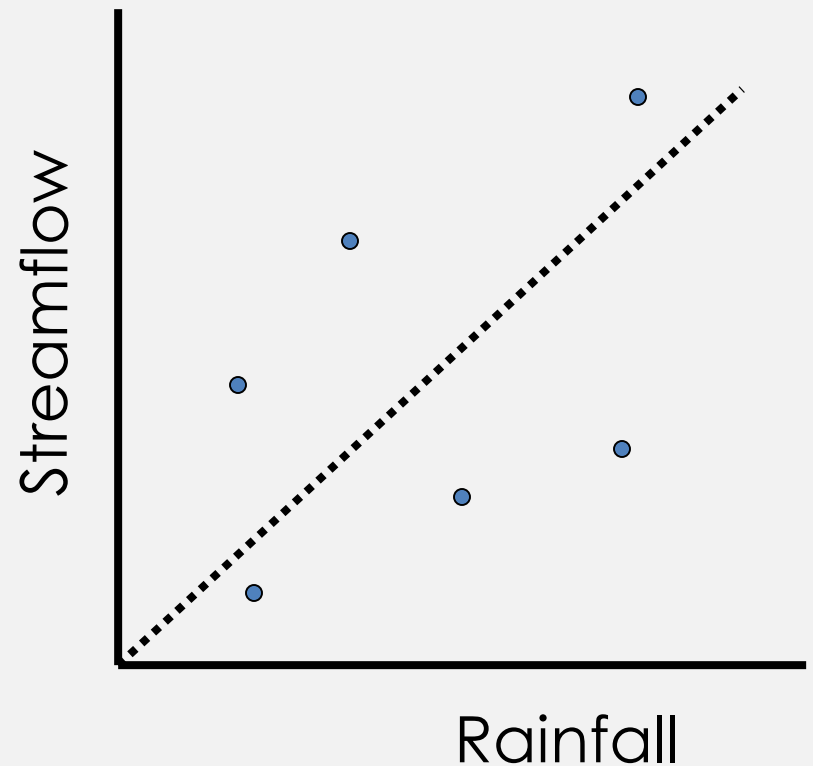
- Bottom-up (from the data) development of an hypothesis
- The hypothesis is generated by the technology directly from the data
- Statistical and machine learning tools such as regression, decision trees and artificial neural networks are used
- Models can be used for *prediction*

Deductive Modeling

- Top-down (toward the data) verification of an hypothesis
- The hypothesis is generated within the mind of the data miner
- Exploratory tools such as data visualization software are used
- Models tend to be used for *description*

Inductive Model

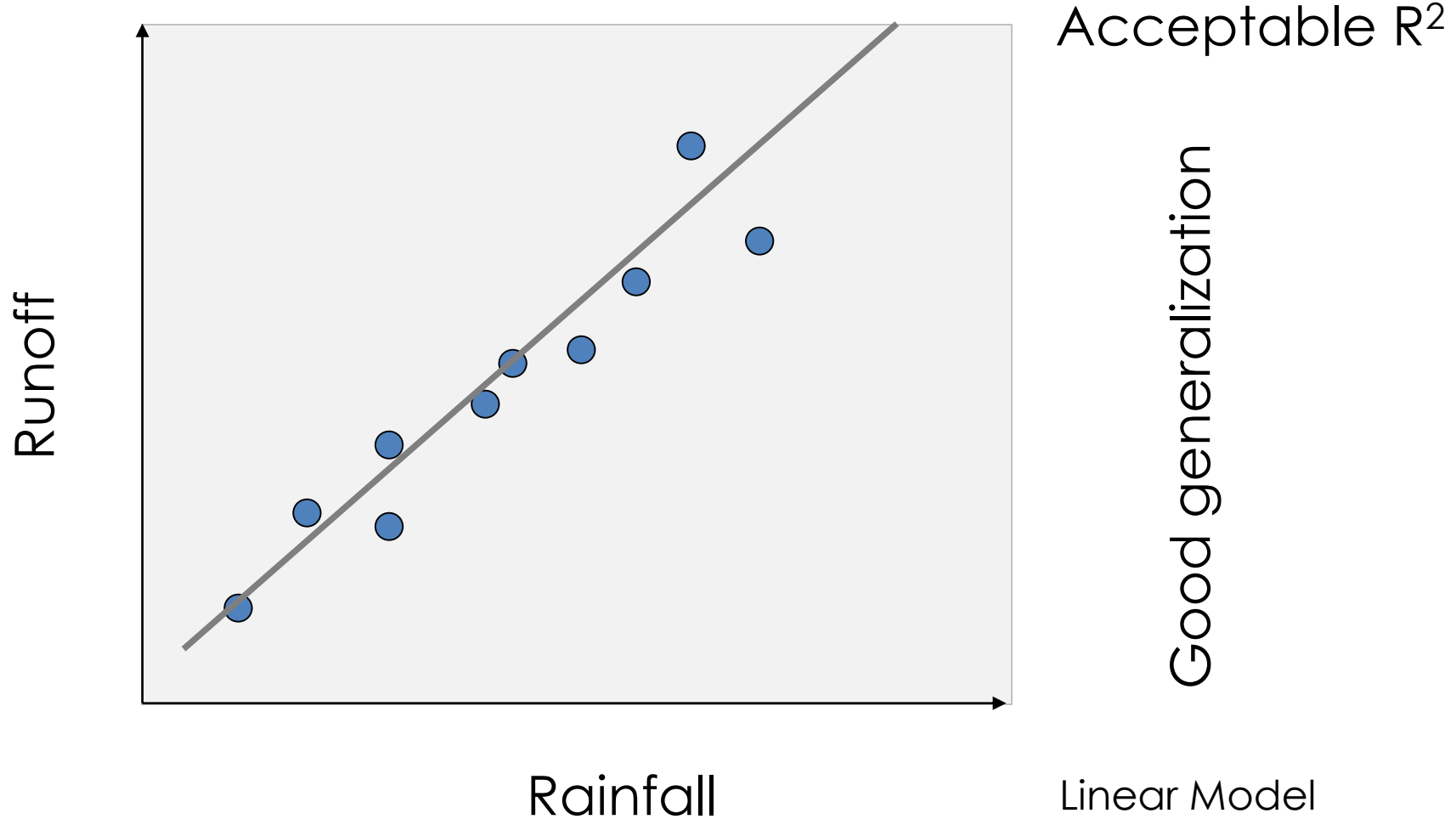
An inductive watershed model can be developed by regressing average streamflow as a function of rainfall.



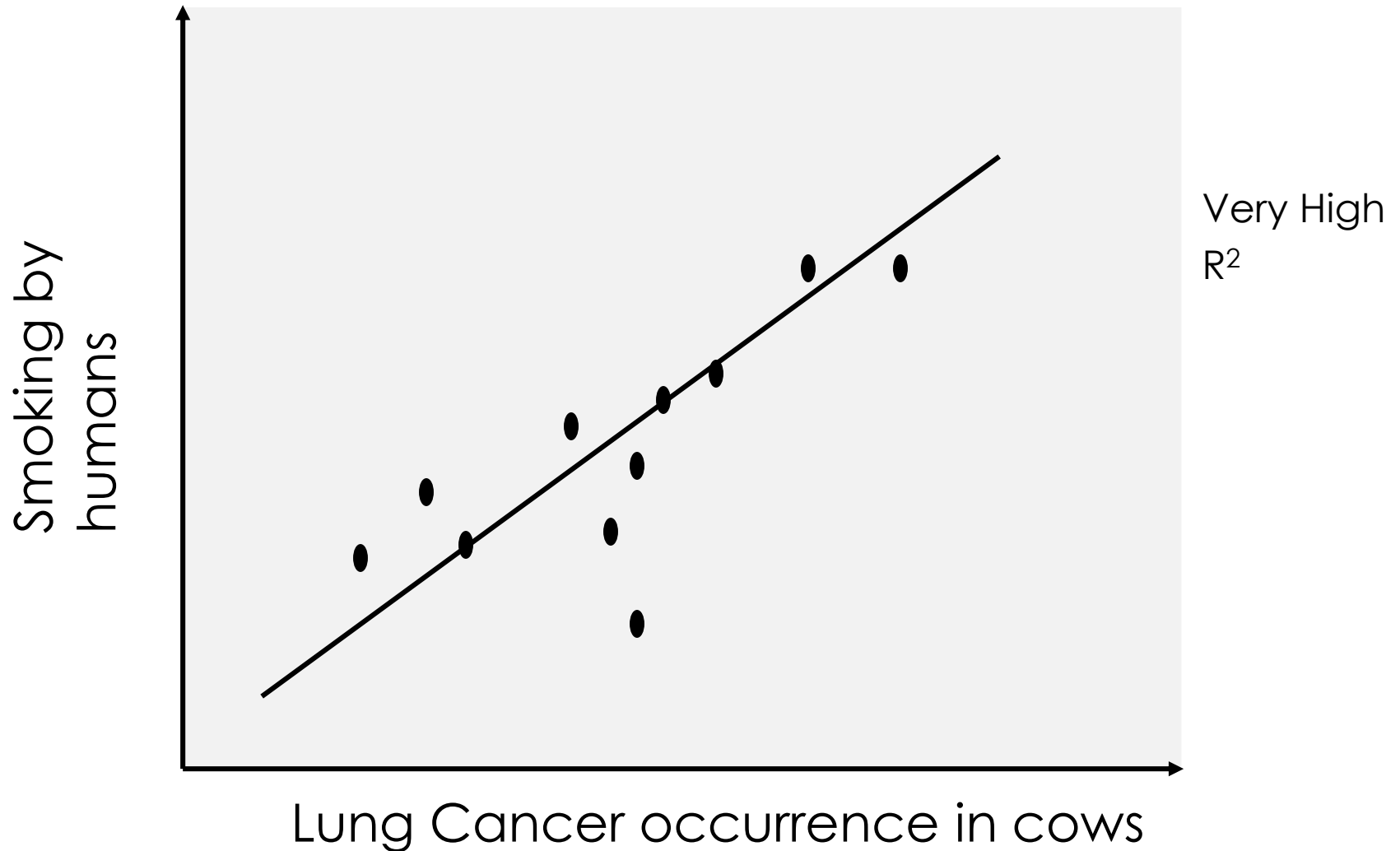
Prescriptive Models

- Prescriptive models analyze the system and provide mathematical representation of the system along with a mechanism to derive optimal outputs for a given system of inputs.

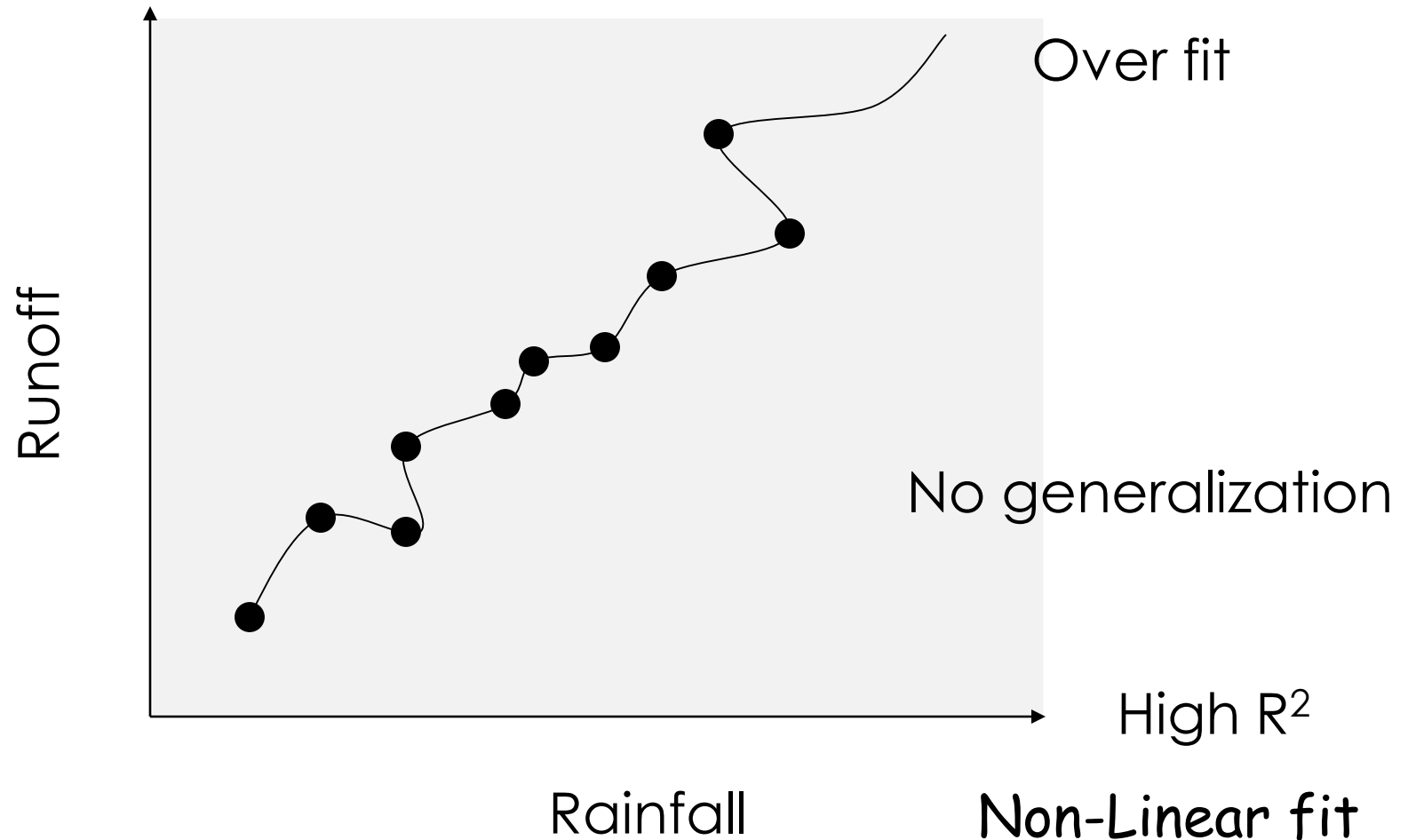
Data-Driven Models



Data-Driven Models: Regression (Spurious Correlation)



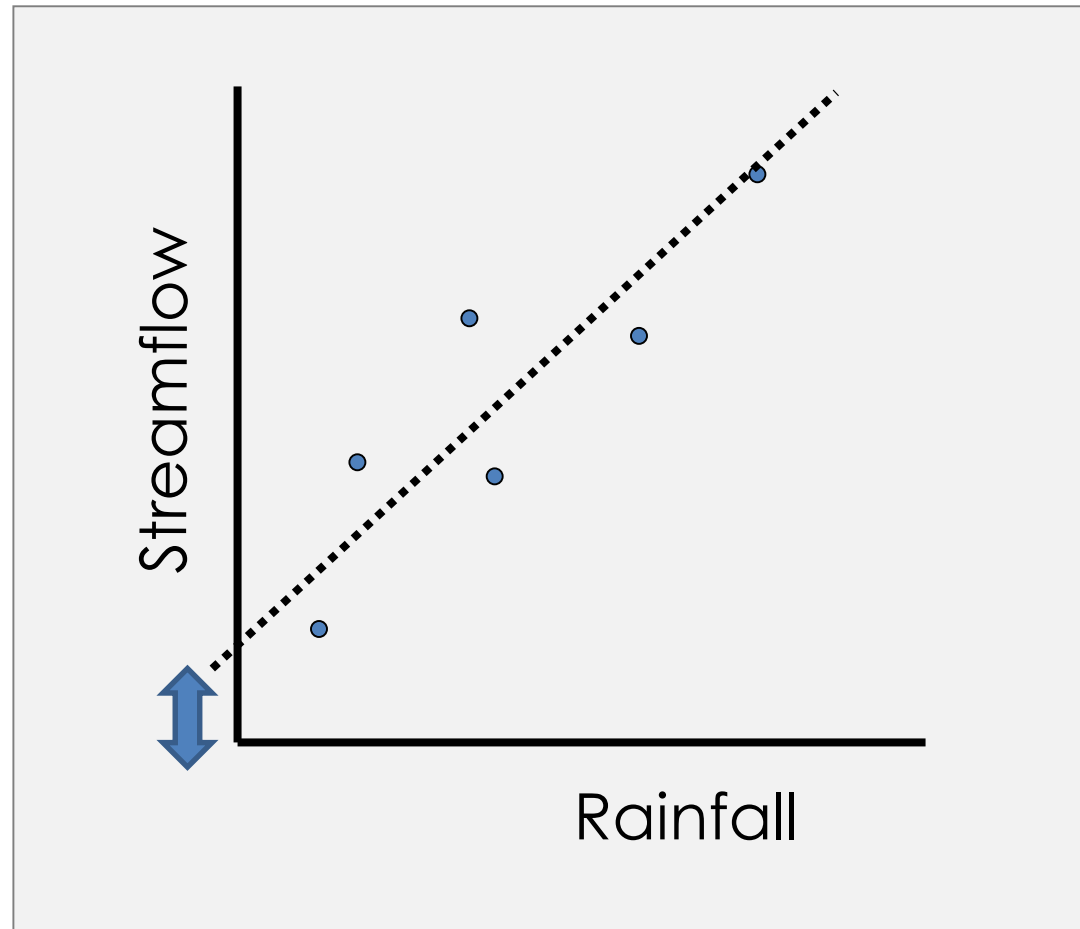
Regression: Over fit



Inductive Models

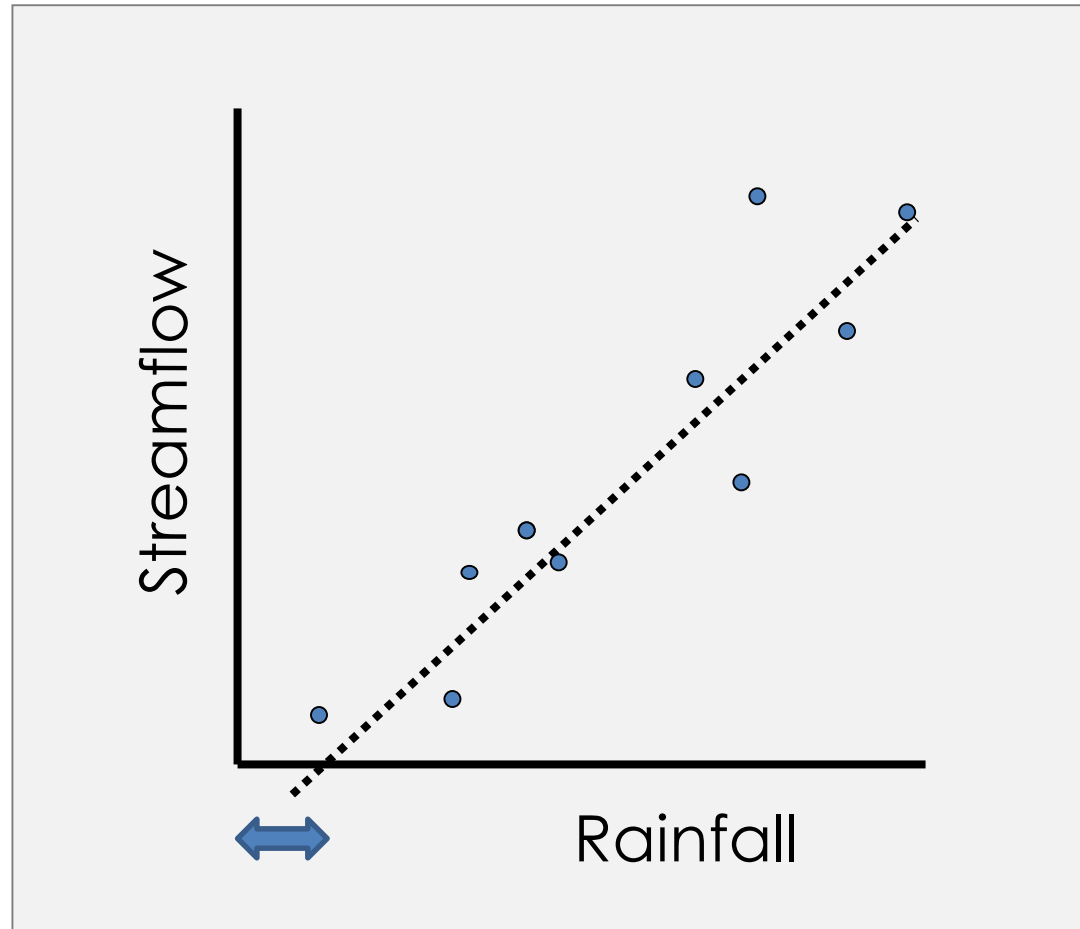
- Use principle of parsimony.
- Use minimum number of parameters of variables in a multi-variate or multi-parameter models
- Aim for good generalization
- High R^2 does not necessarily suggest high generalization.
- Extrapolation beyond the range of data is dangerous and not advisable.
- Use EDA techniques to visually understand the data and evaluate relationships.
- If linear relationship is valid, do not use a spline to fit the relationship.

Understanding processes based on Inductive Models



Baseflow

Understanding processes based on Inductive Models

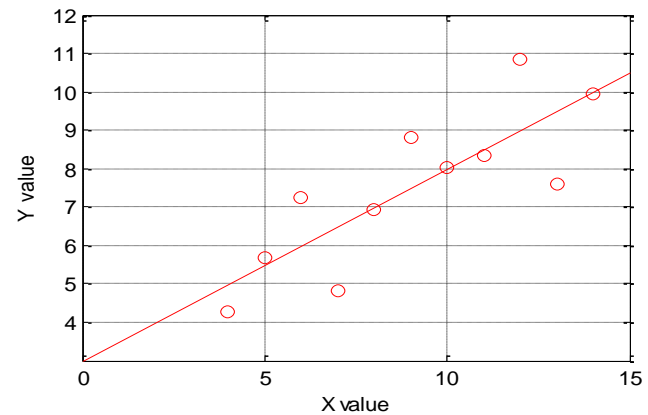
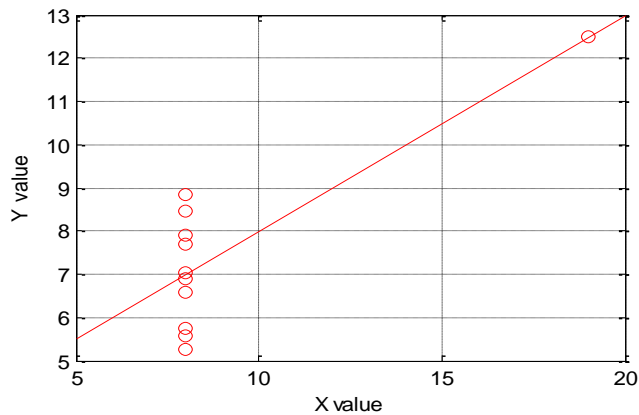
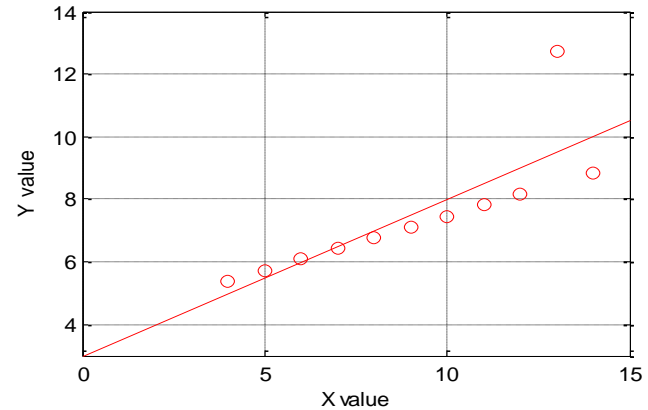
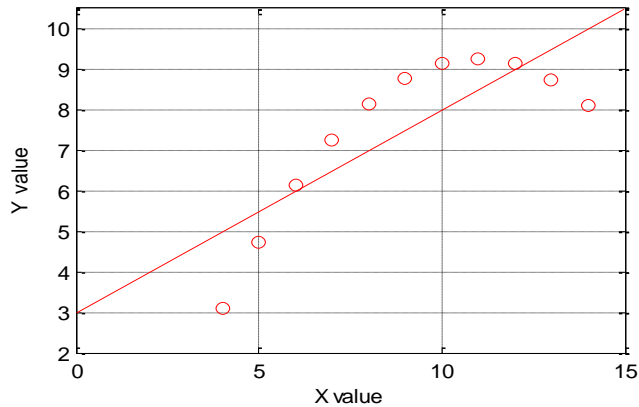


Initial
losses

Anscombe* Datasets

Data set 1		Dataset 2		Dataset 3		Dataset 4	
X	Y	X	Y	X	Y	X	Y
10	9.14	10	7.46	8	6.58	10	8.04
8	8.14	8	6.77	8	5.76	8	6.95
13	8.74	13	12.74	8	7.71	13	7.58
9	8.77	9	7.11	8	8.84	9	8.81
11	9.26	11	7.81	8	8.47	11	8.33
14	8.1	14	8.84	8	7.04	14	9.96
6	6.13	6	6.08	8	5.25	6	7.24
4	3.1	4	5.39	19	12.5	4	4.26
12	9.13	12	8.15	8	5.56	12	10.84
7	7.26	7	6.42	8	7.91	7	4.82
5	4.74	5	5.73	8	6.89	5	5.68

Graphical depiction of data



Anscombe Datasets

- All datasets have similar means
- All datasets have same value of variance or standard deviation.
- The correlation coefficients based two variable datasets are same for all the data

Summary Statistics (Quantitative)

- Minimum
- Maximum
- Mean
- Mode
- Median
- Trimmed Mean
- Standard Deviation and Variance
- Skewness
- Kurtosis
- Percentile
- Range
- Interquartile range
- Variants of Mean – Geomean, Harmonic mean
- Mean Absolute Deviation

Summary Statistics

- Measure(s) of central tendency (location)
- Measure of dispersion (spread)
- Measure of symmetry (Skewness)
- Measure of peakedness (kurtosis)
- Central tendency
 - Mean
 - Mode
 - Median
- Dispersion
 - Variance

Range

- The range (R) provides information about the difference between the maximum and the minimum of a sample dataset.

$$R = \max(\theta_i) - \min(\theta_i)$$



Mean Absolute Deviation

- The mean absolute deviation (MAD) is the mean ($\bar{\theta}$) of deviations of observations from mean of sample dataset.

$$MAD = \frac{1}{N} \sum_{i=1}^N |\theta_i - \bar{\theta}| \quad (1)$$

Median Absolute Deviation (median)

- The mean absolute deviation can also be calculated using the mean of deviation observations from a median value of the sample dataset (θ_M).

$$MAD_M = \frac{1}{N} \sum_{i=1}^N |\theta_i - \theta_M| \quad (1)$$

Skewness and Kurtosis

- Measures of shape are evaluated using skewness coefficient (g) and kurtosis (k) parameters of the dataset. These measures are estimated using the equations 1 and 2 respectively. The parameters are estimated for rain and radar data sets and are compared.

$$g = \frac{\sum_{i=1}^N (\theta_i - \bar{\theta})^3}{(N-1)S^3} \quad (1)$$

$$k = \frac{\sum_{i=1}^N (\theta_i - \bar{\theta})^4}{(N-1)S^4} \quad (2)$$

Interquartile Range (IQR)

- Interquartile range (IQR) refers to the difference between 75th and 25th percentiles of a variable ($Q_3 - Q_1$). The interquartile range is an alternative to the standard deviation and is less affected by extremes than the standard deviation.

$$IQR = Q_3 - Q_1$$

Serial Autocorrelation

- The autocorrelation coefficient is also referred to as serial correlation coefficient. The first-order autocorrelation coefficient can be referred to as correlation coefficient of the first $N-1$ observations, $\theta_1 \dots \dots \theta_{N-1}$, and the next $N-1$ observations, $\theta_2 \dots \dots \theta_N$. These two series are used for calculations of average values as they are referred to as $\bar{\theta}_{(1)}$ and $\bar{\theta}_{(2)}$ respectively. The autocorrelation values can be obtained for different lag (t) values as given by equation

$$\rho_t = \frac{\sum_{i=1}^{N-t} (\theta_i - \bar{\theta}_{(1)}) (\theta_{i+t} - \bar{\theta}_{(2)})}{\sqrt{\sum_{i=1}^{N-t} (\theta_i - \bar{\theta}_{(1)})^2} \sqrt{\sum_{i=2}^N (\theta_i - \bar{\theta}_{(2)})^2}}$$

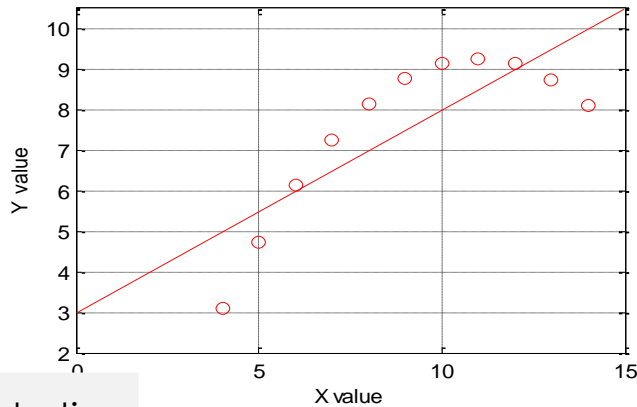
EDA – Graphical techniques

- Graphical techniques employed in EDA are often quite simple, consisting of various techniques such as:
 - Plotting the raw data (such as data traces, histograms, [bi-histograms](#), probability plots, lag plots and [Youden](#) plots).
 - Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
 - Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

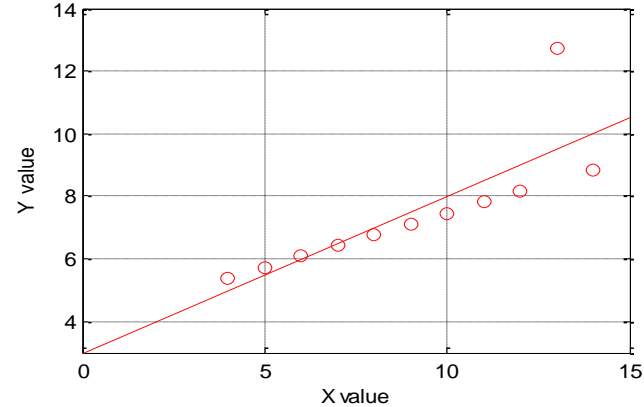
Anscombe Datasets

Data set 1		Dataset 2		Dataset 3		Dataset 4	
X	Y	X	Y	X	Y	X	Y
10	9.14	10	7.46	8	6.58	10	8.04
8	8.14	8	6.77	8	5.76	8	6.95
13	8.74	13	12.74	8	7.71	13	7.58
9	8.77	9	7.11	8	8.84	9	8.81
11	9.26	11	7.81	8	8.47	11	8.33
14	8.1	14	8.84	8	7.04	14	9.96
6	6.13	6	6.08	8	5.25	6	7.24
4	3.1	4	5.39	19	12.5	4	4.26
12	9.13	12	8.15	8	5.56	12	10.84
7	7.26	7	6.42	8	7.91	7	4.82
5	4.74	5	5.73	8	6.89	5	5.68

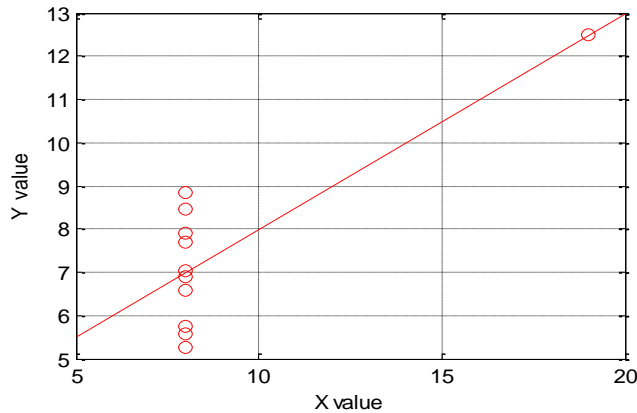
Graphical depiction of data



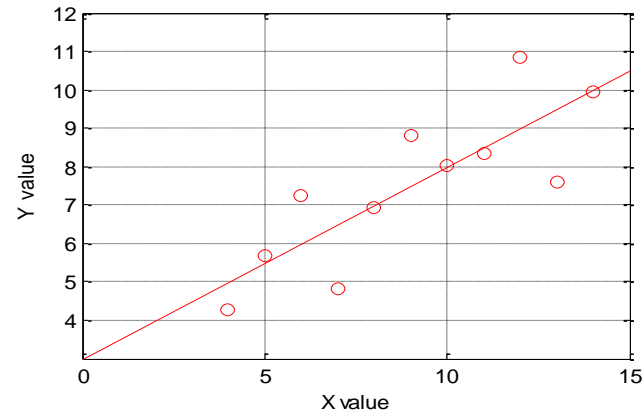
Quadratic.



Outlier.



Problems with data



Linear with some scatter.

Visual Exploration

Uni-variate dataset

- Run-Sequence plot (or time series plot)
- Boxplot
- Histogram
- Autocorrelation
- Lag Plot
- Stem and Leaf Plot
- Bar Graph
- Line Graph
- Probability Density Function (parametric)
- Cumulative Distribution Function (non-parametric)
- Kernel Density Estimate (non-parametric)

Bi-variate datasets

- Bi-histogram (two histograms together – one below the other)
- Scatter plot
- Scatter histogram

Multivariate dataset (3-D)

- Contour plot

Examples of Graphical Analysis (univariate data)

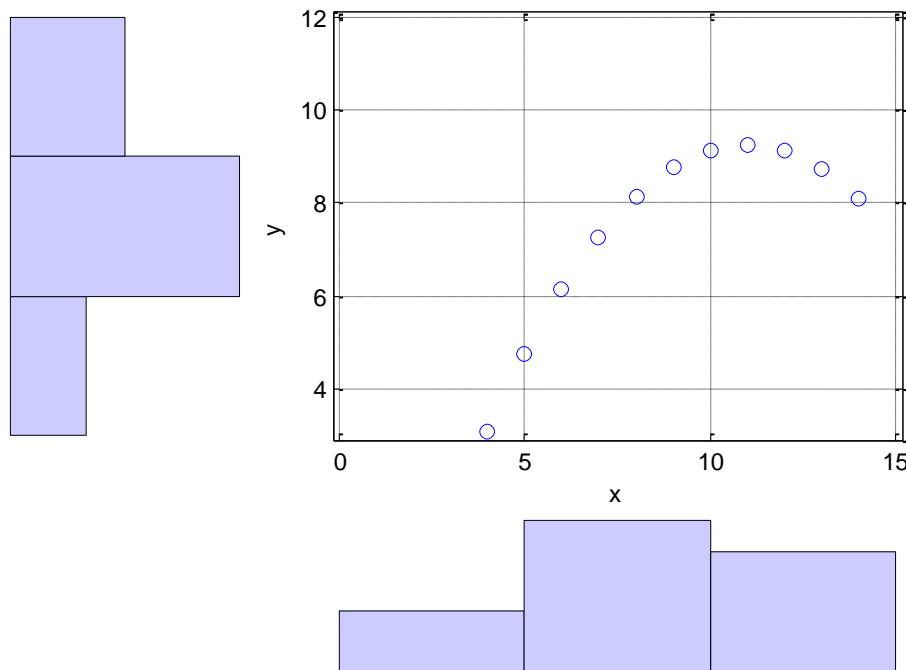
- Run Sequence plot
 - A time series plot (plot of variable value with respect to time).
- Box Plot
 - A plot which summarizes the data showing median, inter-quartile range, 25th percentile, 75th percentile.
 - Help understand the distribution of the data.
 - Compare the medians if several box plots are plotted using multiple datasets.
 - Assess outliers.

Examples of Graphical Analysis (univariate data)

- Histogram
 - Helps to assess the distribution of the data.
- Probability Density Function (PDF)
 - Help to assess the distribution of both discrete and continuous datasets.
- Cumulative Distribution Function (CDF)
 - Helps understand the data by providing non-exceedance probabilities.

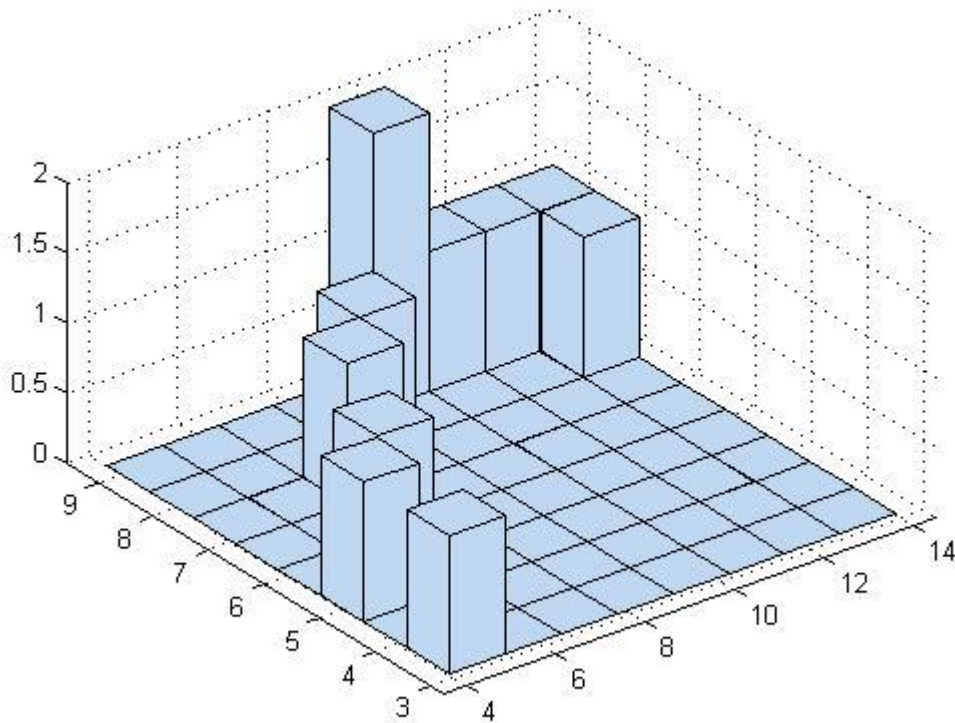
Scatter histogram

Dataset # 1

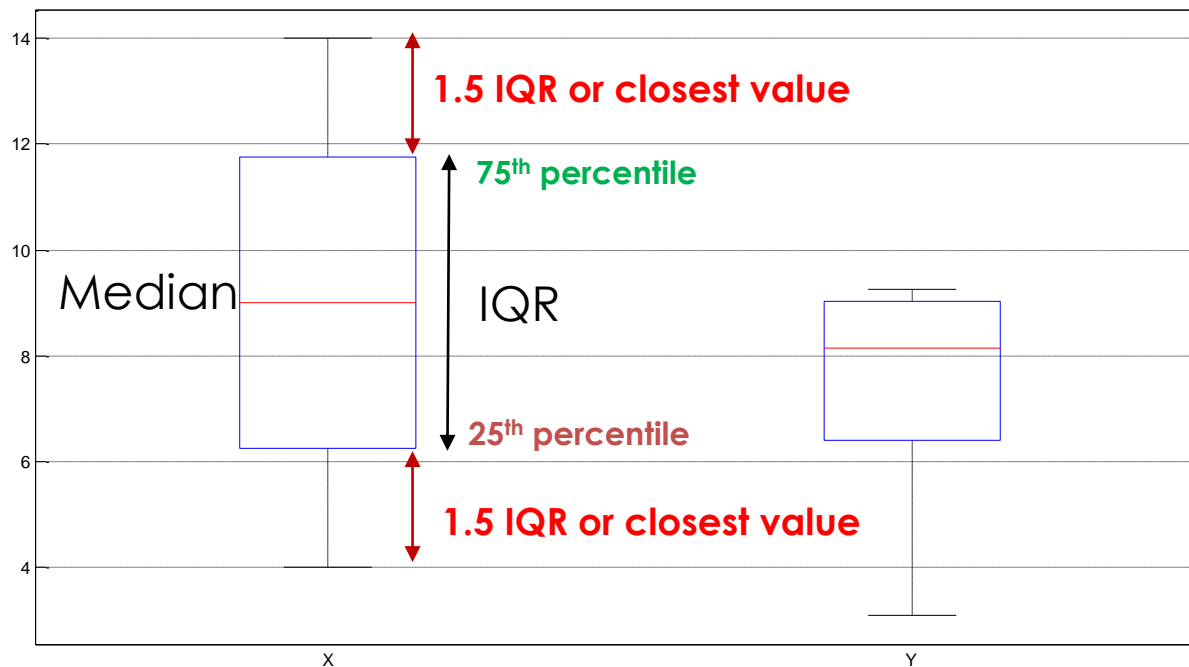


3-D histogram

Dataset # 1



Boxplot

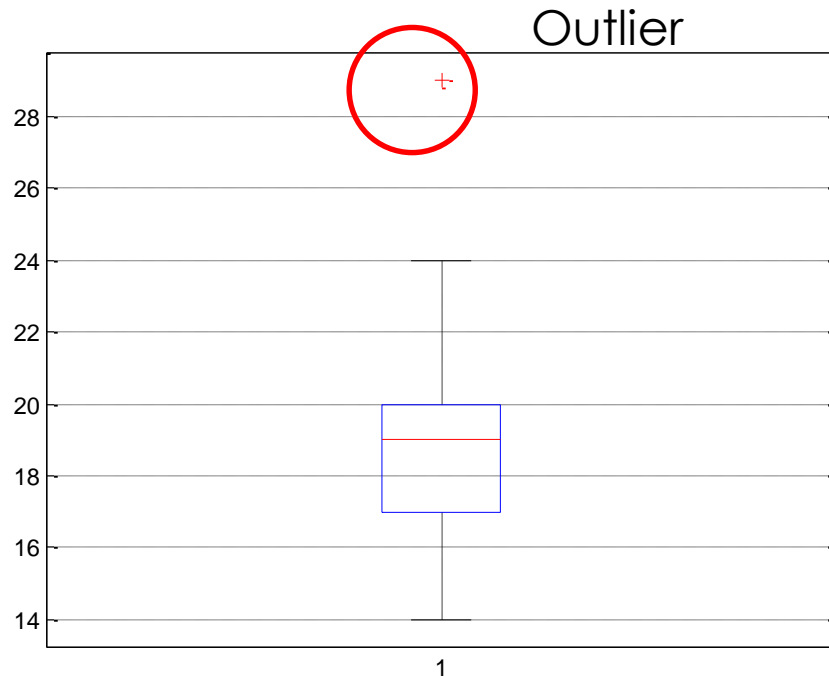


75th percentile value = 11.75
25th percentile value = 6.25
IQR = 5.5
1.5 * IQR = 8.25

Upper Inner Fence = $11.75 + 8.25 = 20$
Lower Inner Fence = $6.25 - 8.25 = -2$

Dataset # 1

Another Example of Boxplot



Data

14
15
16
16
17
17
17
17
17
17
18
18
18
18
18
18
18
18
18
19
19
19
19
20
20
20
20
20
20
20
20
20
21
21
22
23
24
24
29

75th percentile value = 20
25th percentile value = 17
IQR = 3
 $1.5 * \text{IQR} = 4.5$

Upper Inner Fence = $20 + 4.5 = 24.5$ \Rightarrow 24
Lower Inner Fence = $17 - 4.5 = 12.5$ \Rightarrow 14
Closest Values from data

Outlier --- more than 1.5IQR
Extreme Outlier - more than $3*\text{IQR}$

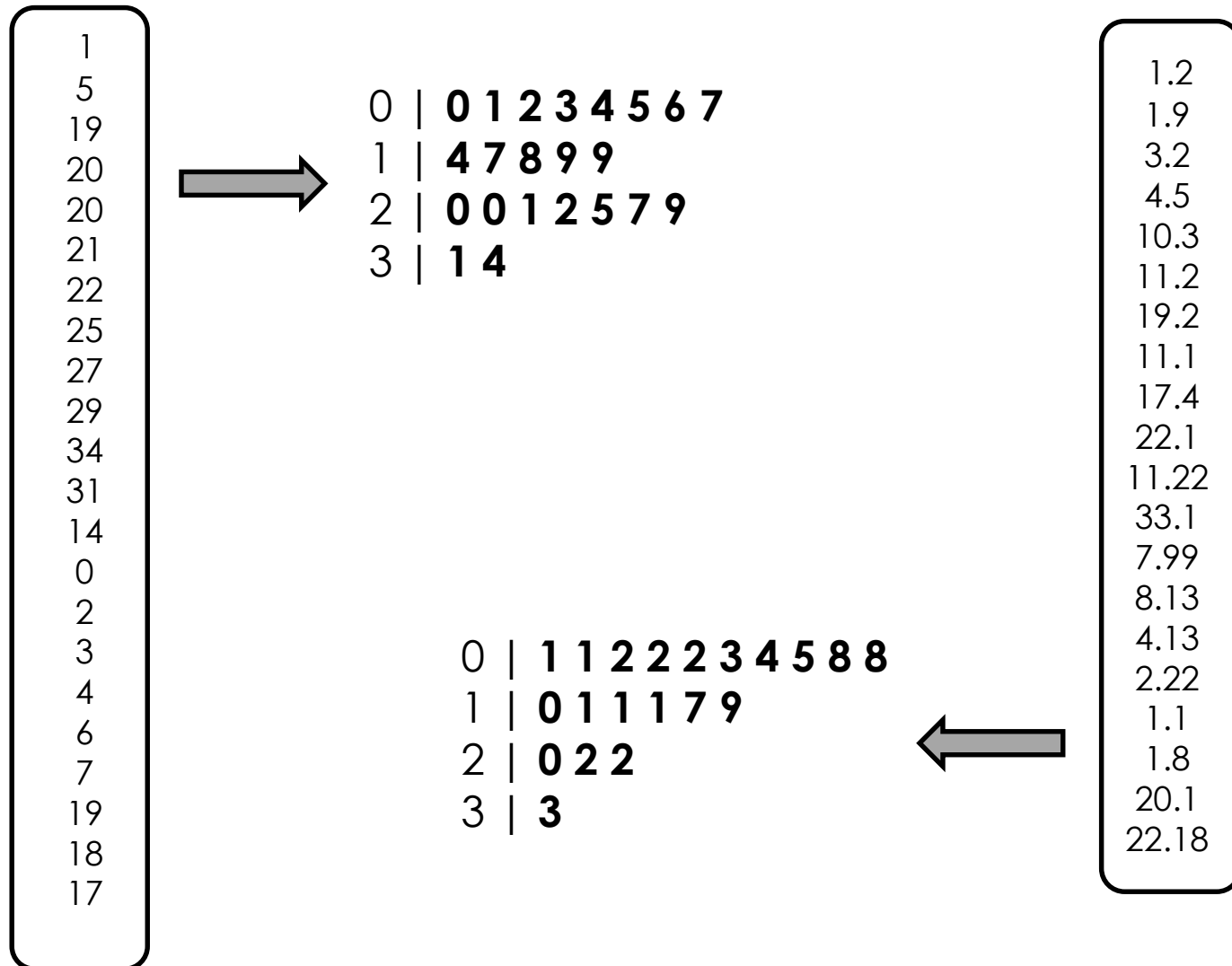
Outliers

- An outlier is an observation that appears to deviate markedly from other observations in the sample.
- Identification of potential outliers is important
- An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).
- In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting.
- In any event, we typically do not want to simply delete the outlying observation. However, if the data contains significant outliers, we may need to consider the use of robust statistical techniques.

Handling Outliers

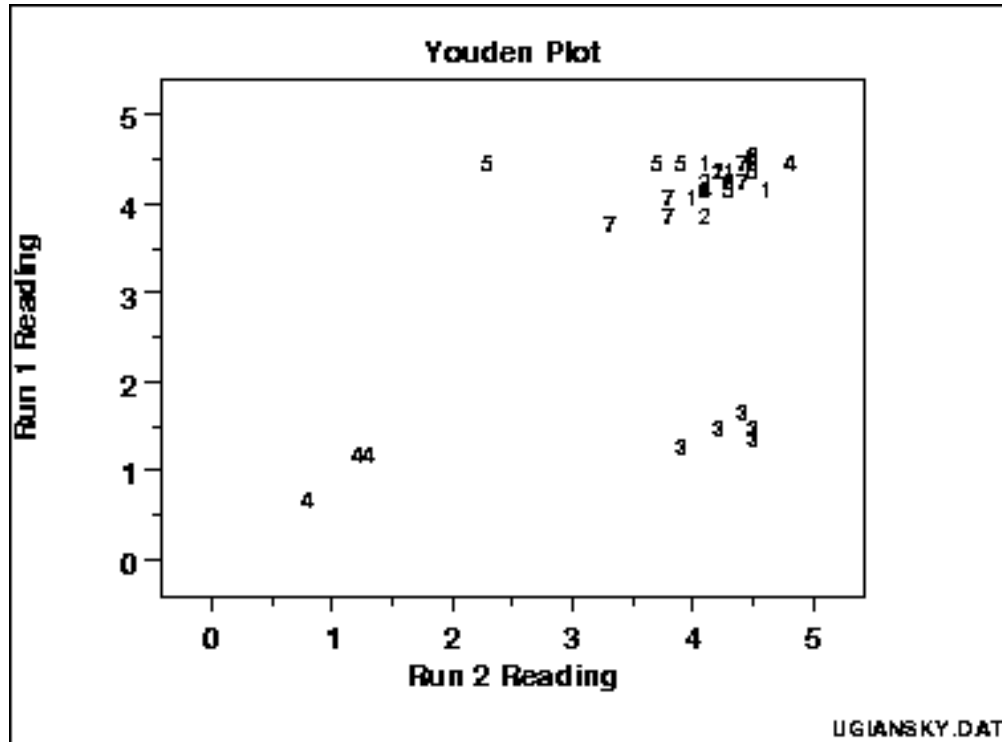
- Outlier labeling - flag potential outliers for further investigation (i.e., are the potential outliers erroneous data, indicative of an inappropriate distributional model, and so on).
- Outlier accommodation - use robust statistical techniques that will not be unduly affected by outliers. That is, if we cannot determine that potential outliers are erroneous observations, do we need modify our statistical analysis to more appropriately account for these observations?
- Outlier identification - formally test whether observations are outliers.
- Outliers – “Distributional Monsters”

Stem and Leaf Plot



Youden Plot

Example



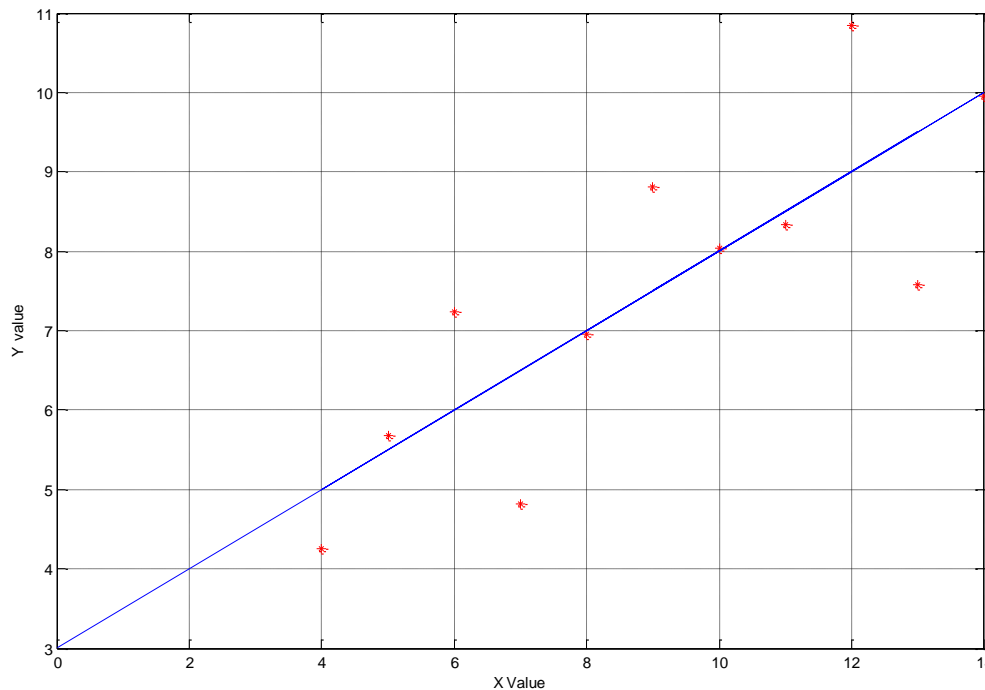
Example: Youden plots
Useful for analyzing inter-lab data when each lab has made two runs on the same product or one run on two different products.

Not all labs are equivalent.
Lab 4 is biased low.
Lab 3 has within-lab variability problems.
Lab 5 has an outlying run.

Are measurements coming from different laboratories equivalent?

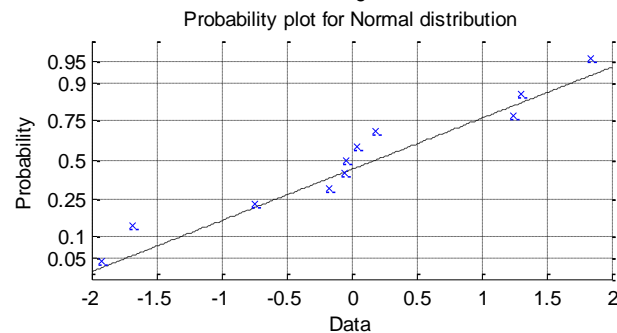
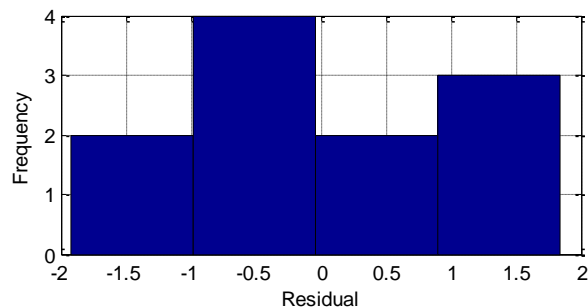
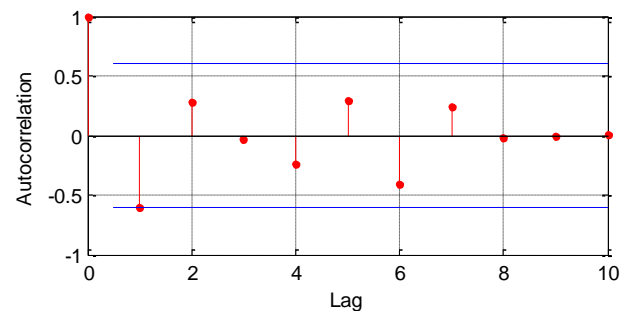
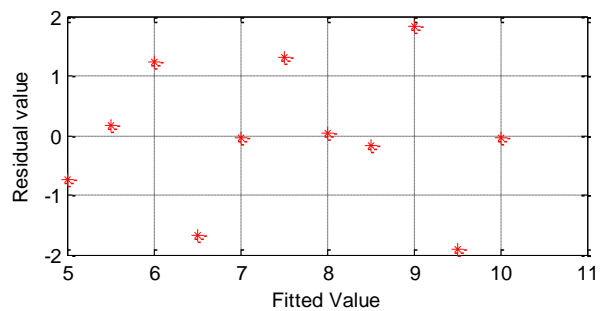
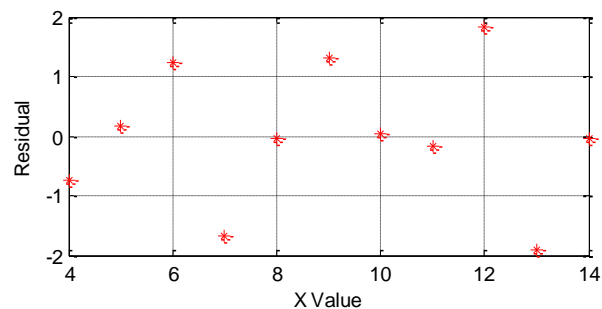
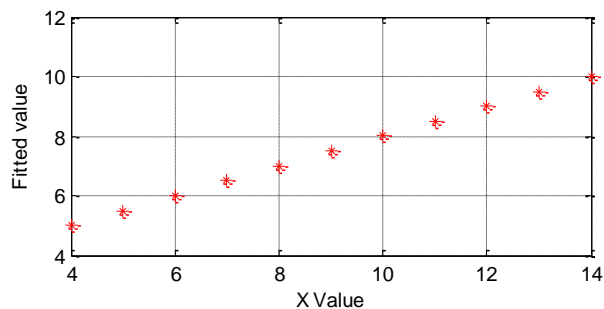
Regression

- Linear regression Model ($Y = mX + c$)



$M=0.5001$
 $C=3.001$

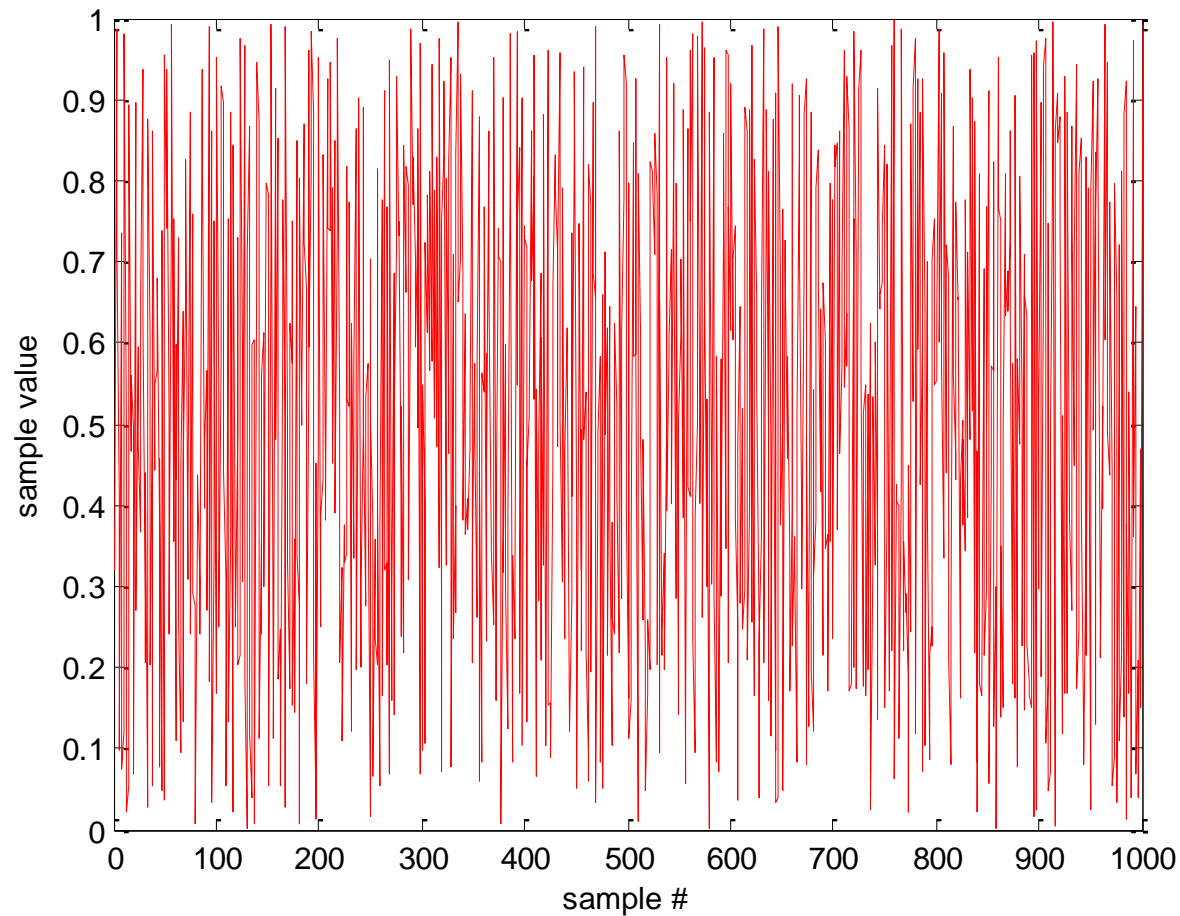
Evaluation of the Model



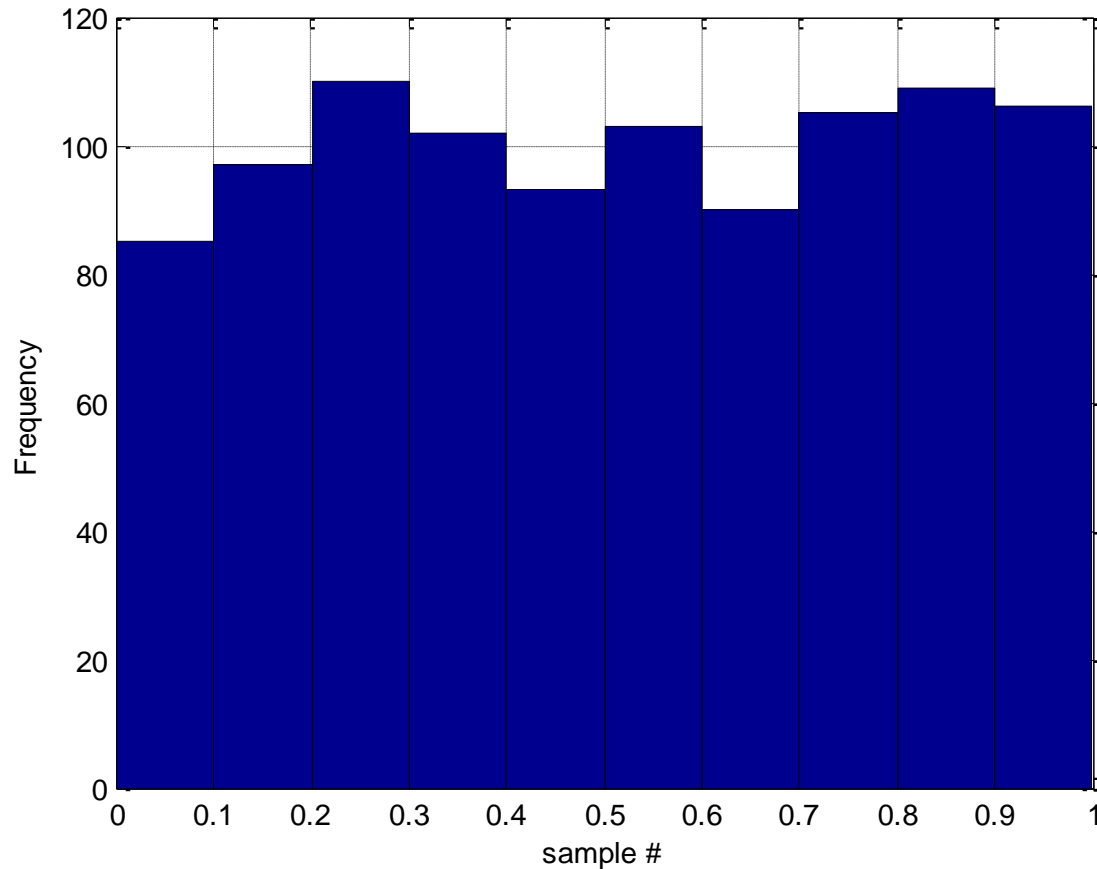
EDA to understand data

- 1000 uniform random numbers between 0 and 1. Our sample dataset.
- A set of uniform random numbers are used to illustrate the effects of a known underlying non-normal distribution.
- Check the properties of the data
- Evaluate possible fit (distributional)
- Confirm randomness

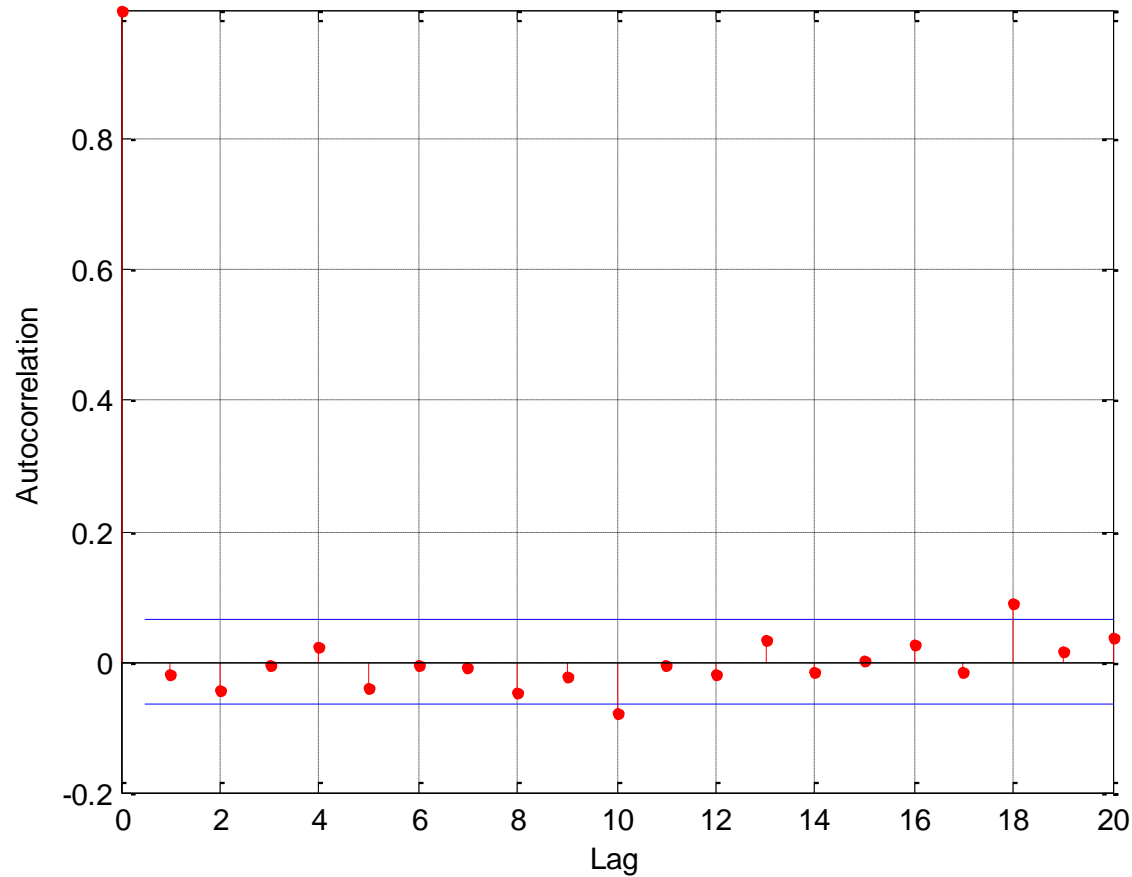
Run-Sequence Plot



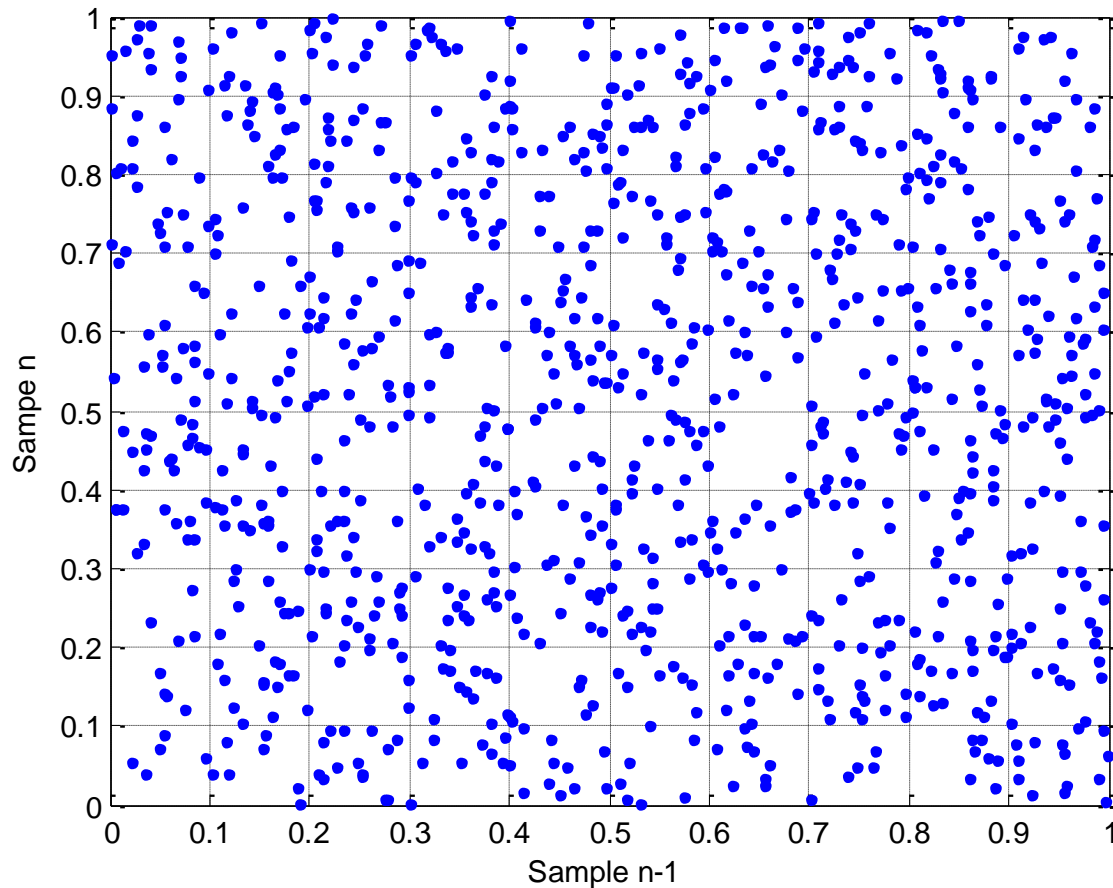
Histogram of data



Autocorrelation plot



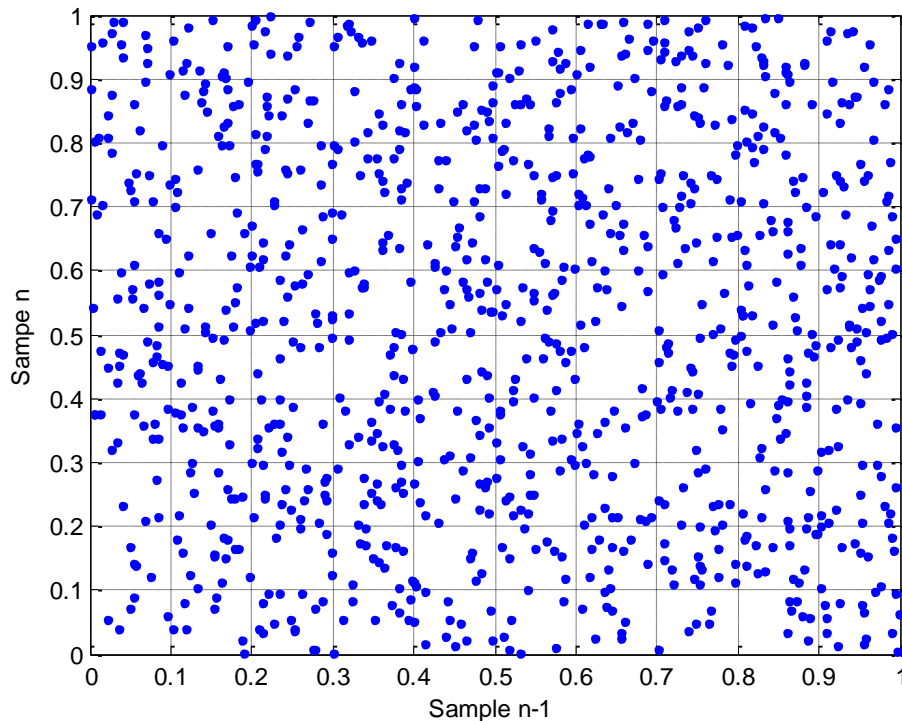
Lag Plot



Looks completely Random.

No relationship between any two consecutive sample series

Randomness

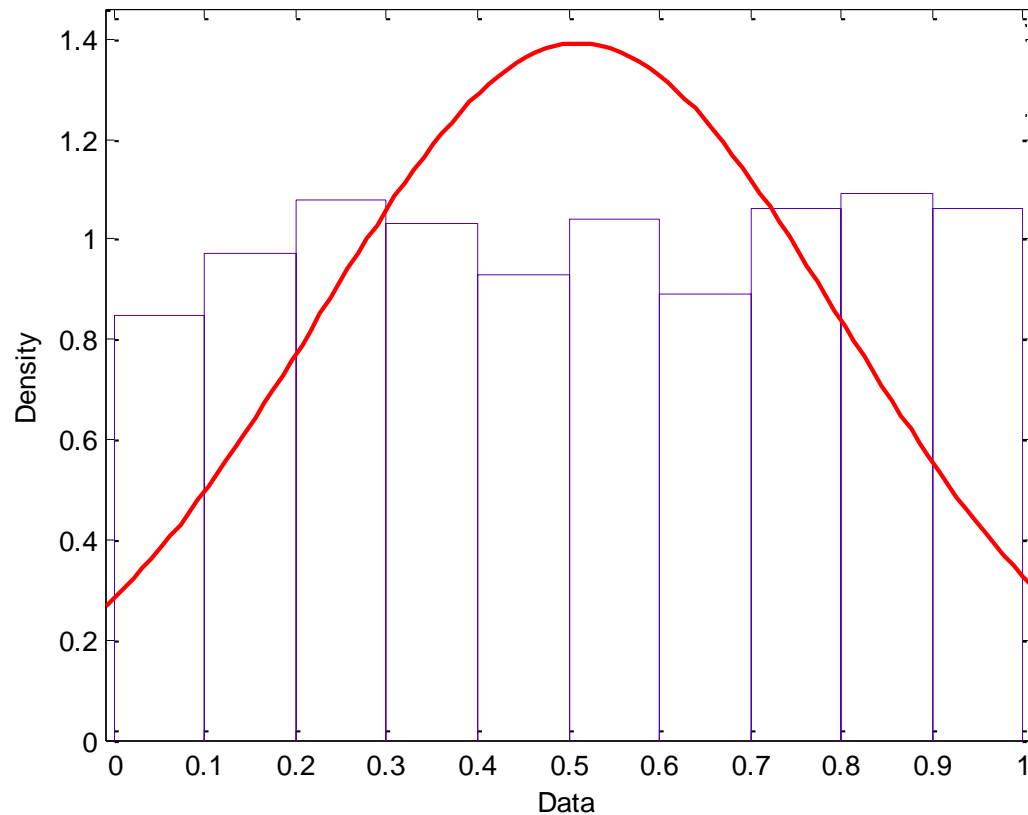


- The data are random.
- The data exhibit no autocorrelation.
- The data contain no outliers.

The lag plot is for lag = 1.
Absence of any structure is evident.
One cannot infer, from a current value \mathbf{S}_{n-1} , the next value \mathbf{S}_n .

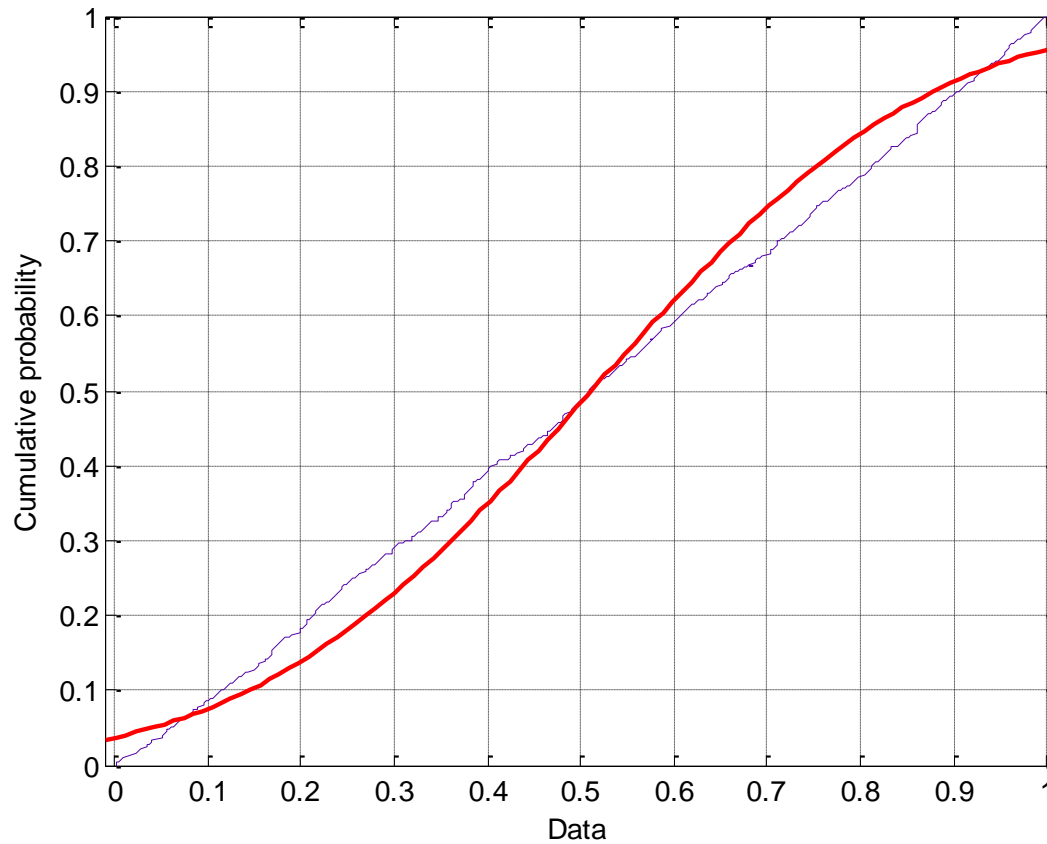
Thus for a known value \mathbf{S}_{n-1} on the horizontal axis (say, $\mathbf{S}_{n-1} = 0.8$), the \mathbf{S}_n -th value could be virtually anything (from $\mathbf{S}_n = 0.1$ to $\mathbf{S}_n = 1.0$). Such non-association is the essence of randomness.

Fitting Normal distribution

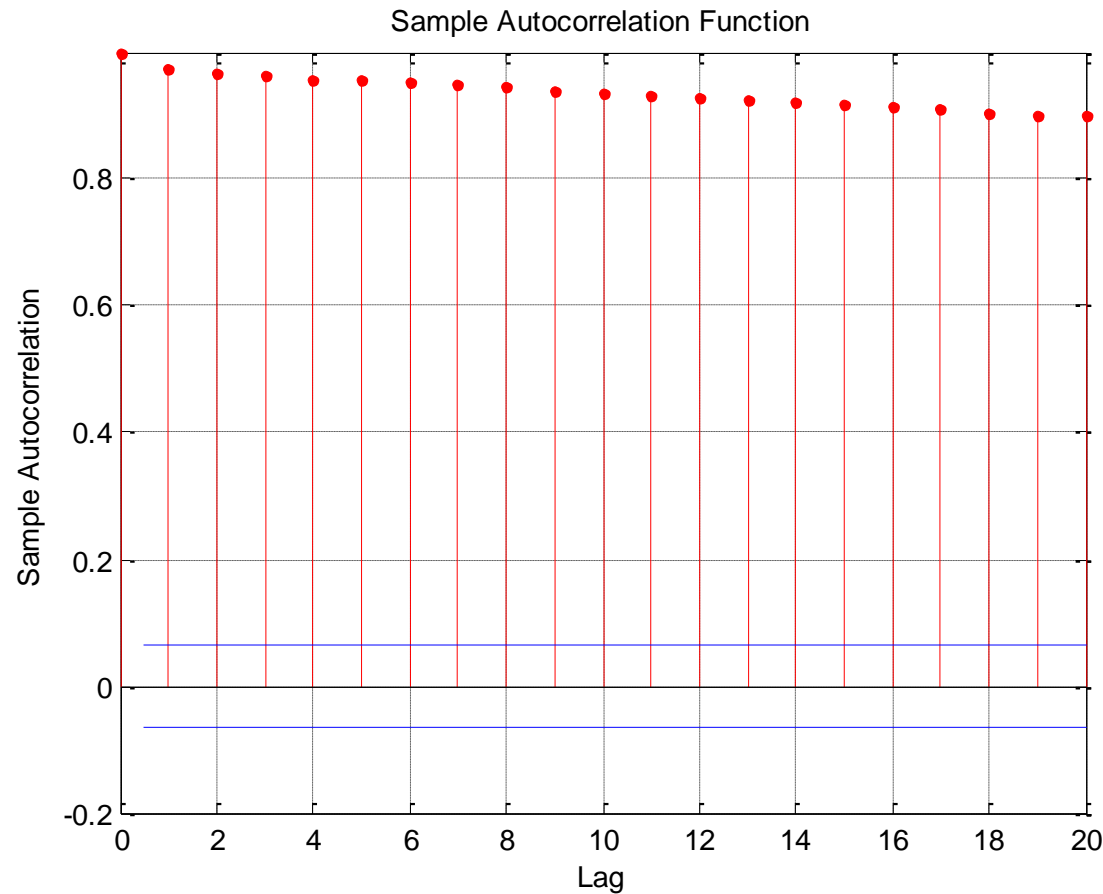


Normal Distribution is not a good fit

Fit shown by CDF



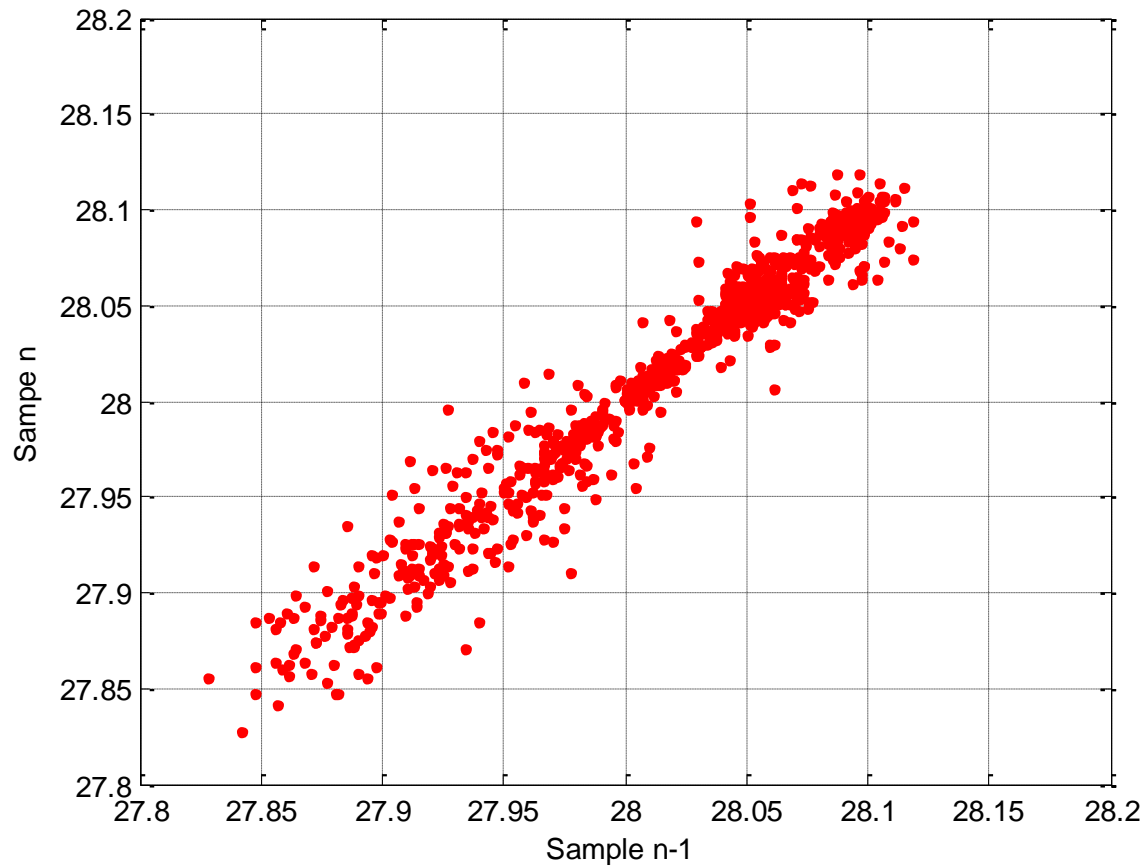
High Autocorrelation



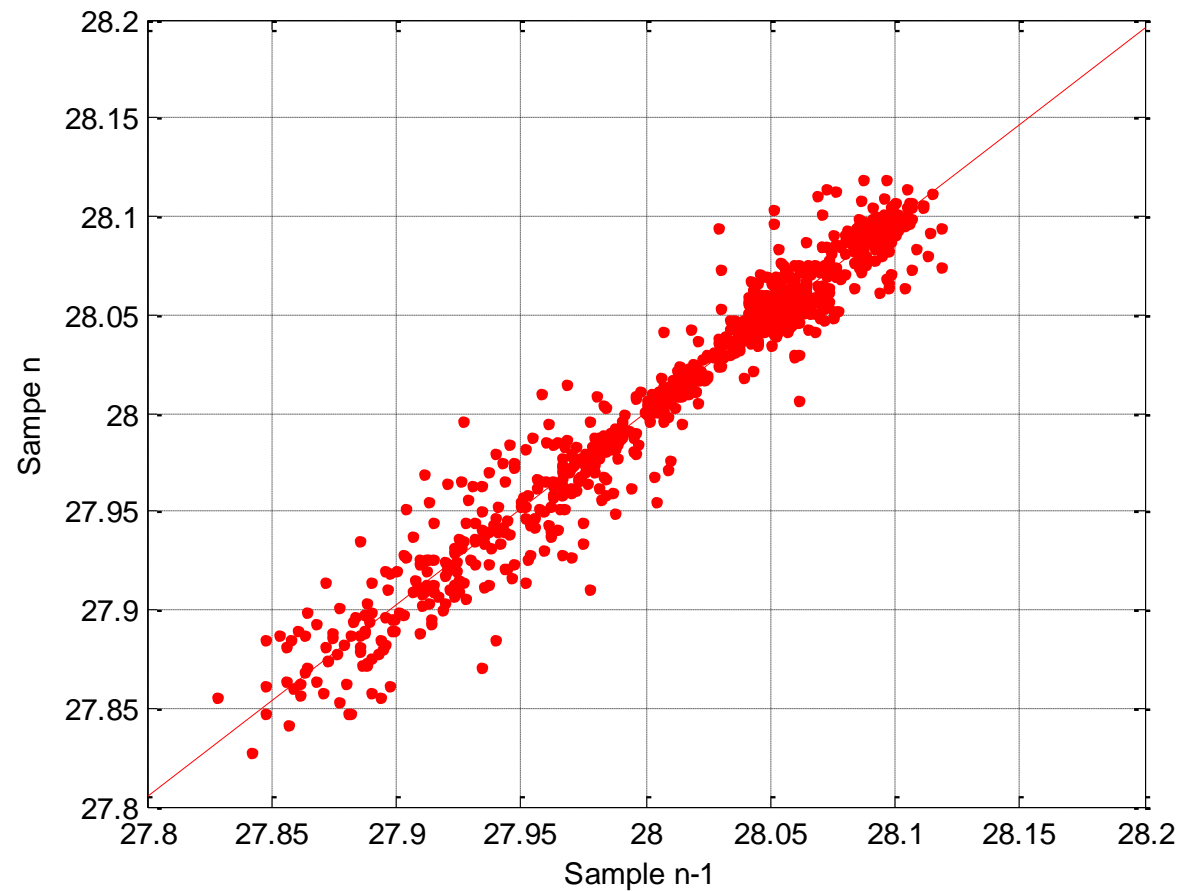
Very high
autocorrelation at
different lags

Data based on observed resistor values, NIST

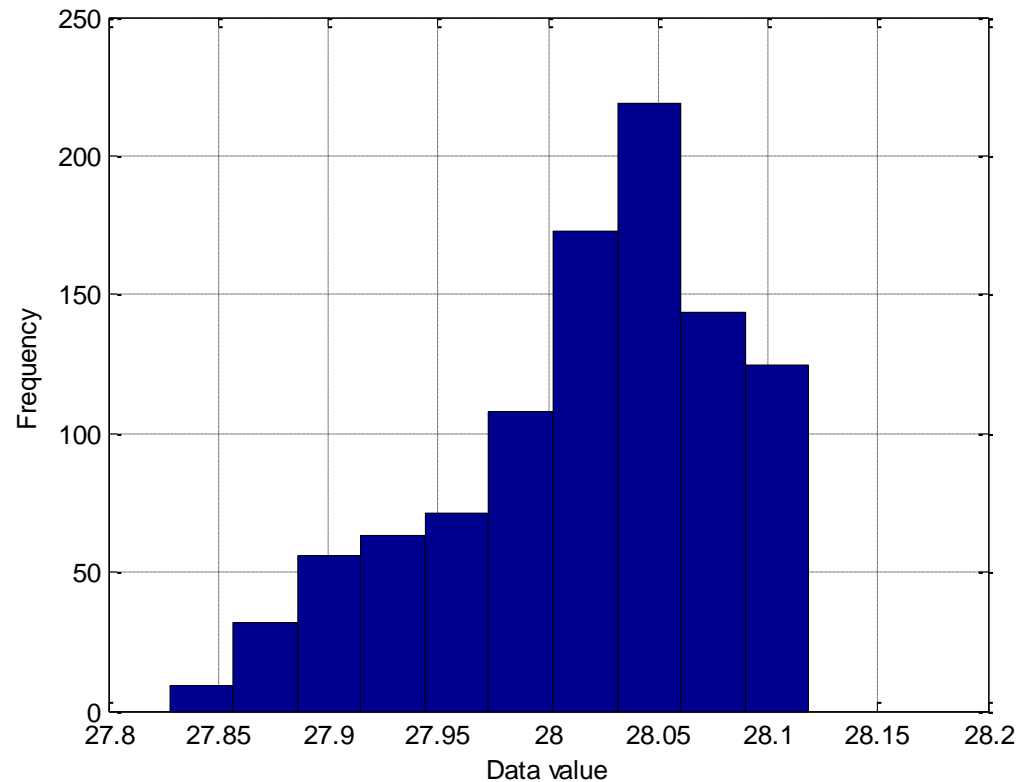
Lag Plot



Prediction Model



Distribution of data



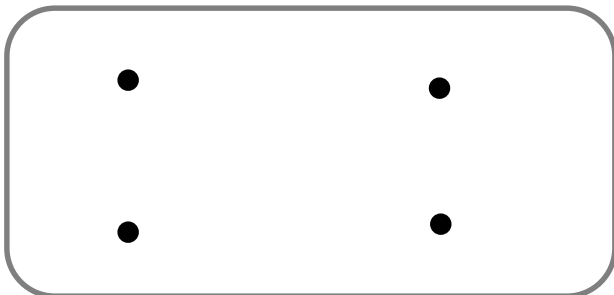
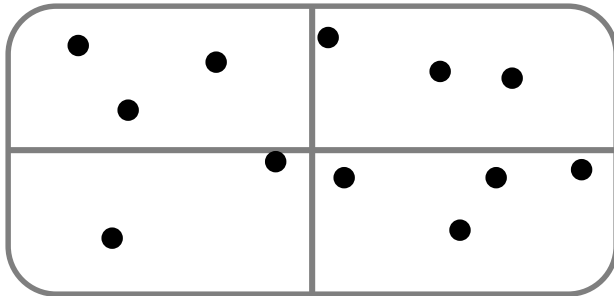
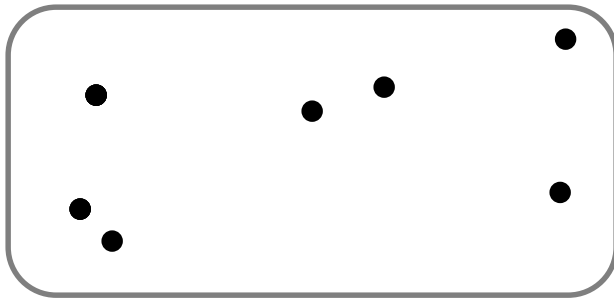
Overview of Spatial Analysis

- To understand a process or a phenomenon in space, sampling of the variables that affect the process is carried out.
- Sampling requires
 - Good knowledge of the variable in question
 - Objectives of sampling are clear
 - Economics are considered
 - Logistics of collection, transport and evaluation of observations are considered.

Observations in Space

- Development of Sampling schemes
- Sampling Schemes are developed for observation/monitoring/estimation of a number of **hydroclimatic** and **environmental variables**.
- Generally sampling is done at a point level (i.e. a point in space). However, observed available from different sampling schemes can be converted to any **spatial** and temporal resolution.

Data Sampling in Space



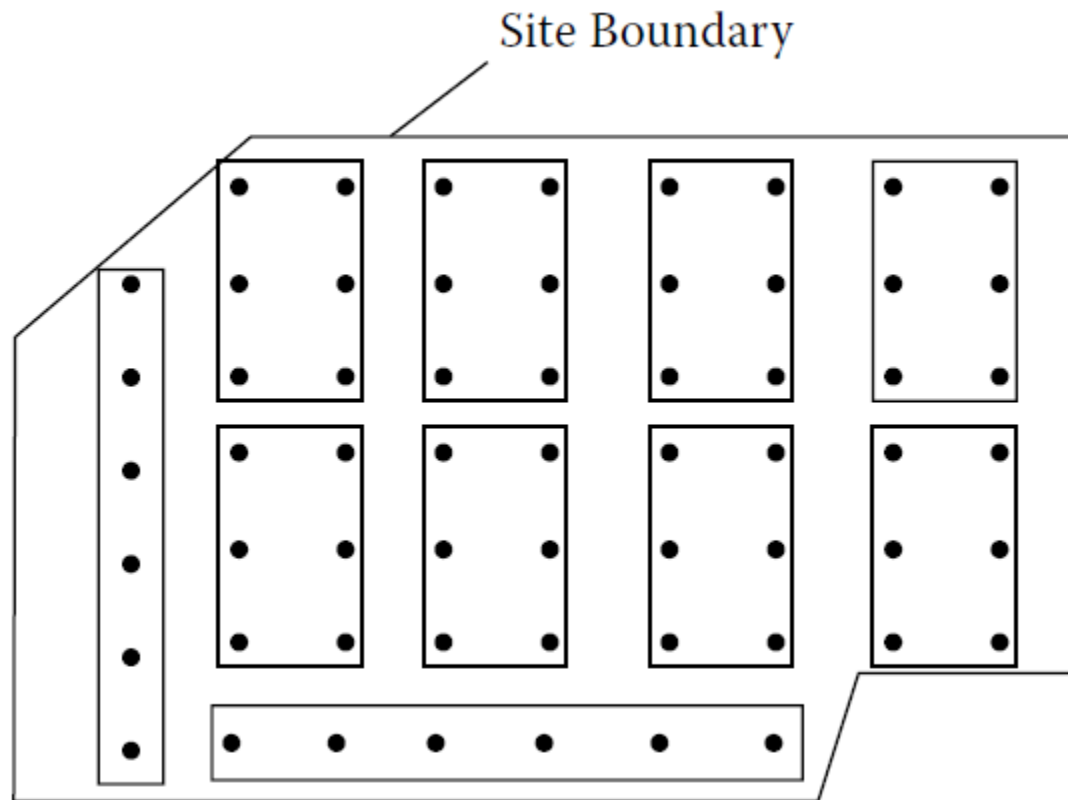
Random Sampling

Stratified Random
Sampling

Systematic Sampling
(non random)

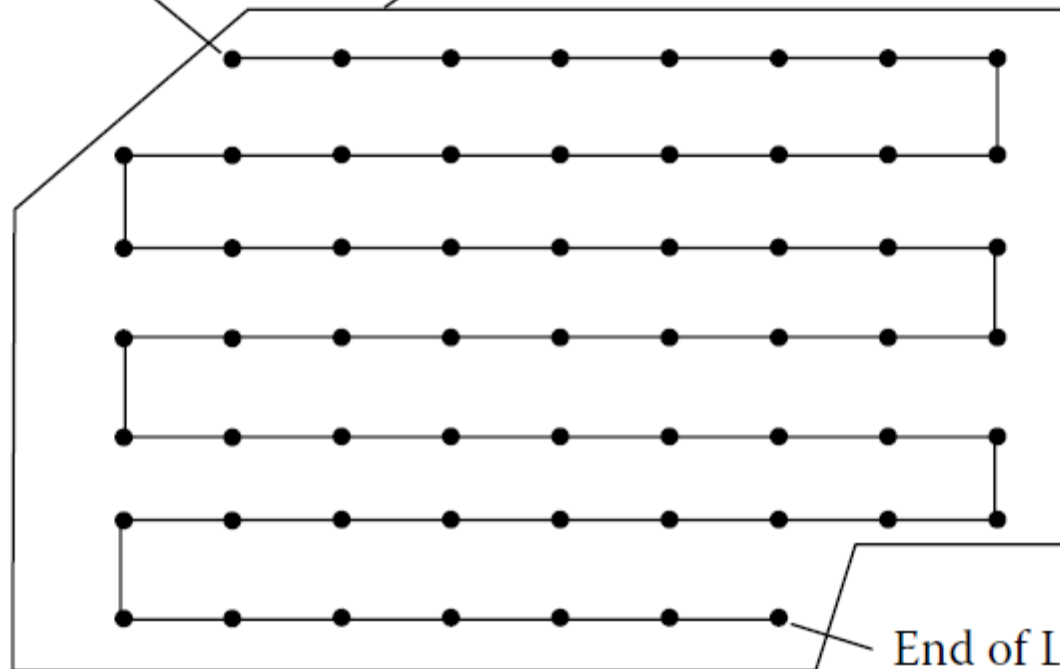
Issues with Random Sampling

- Simple random sampling may introduce clustering. This is problematic when used for use of data for analysis.
 - **Clustering** reduces the spatial sampling variability
 - Increases the bias in your analysis towards one particular results
 - Analysis obtained is not representative of the actual variation of the data variable
 - Conclusions drawn from the data are not valid and generalization can be carried out about a process.



Start of Line
(point 1)

Site Boundary



End of Line
(point 60)

Objectives of Sampling

- A crucial early task in any sampling study is to define the **population of interest** and the sample units that make up this population. This may or may not be straightforward.
- **Simple random sampling** (SRS) involves choosing sample units in such a way that each unit is equally likely to be selected. It can be carried out with or without replacement.

Stratified Sampling

- Stratified random sampling is sometimes useful for ensuring that the units that are measured are well representative of the population.
- However, there are potential problems due to the use of incorrect criteria for stratification or the wish to use a different form of stratification to analyze the data after they have been collected.

Systematic Sampling

- With systematic sampling, the units measured consist of every k th item in a list, or are regularly spaced over the study area.
- Units may be easier to select than they are with random sampling, and estimates may be more precise than they would otherwise be because they represent the whole population well.
- Treating a systematic sample as a simple random sample may overestimate the true level of sampling error.

Issues with Random Sampling

- Simple random sampling may introduce clustering. This is problematic when used for use of data for analysis.
 - **Clustering** reduces the spatial sampling variability
 - Increases the bias in your analysis towards one particular results
 - Analysis obtained is not representative of the actual variation of the data variable
 - Conclusions drawn from the data are not valid and generalization can be carried out about a process.

Objectives of Sampling

- A crucial early task in any sampling study is to define the population of interest and the sample units that make up this population. This may or may not be straightforward.
- Simple random sampling (SRS) involves choosing sample units in such a way that each unit is equally likely to be selected. It can be carried out with or without replacement.

Stratified Sampling

- Stratified random sampling is sometimes useful for ensuring that the units that are measured are well representative of the population.
- However, there are potential problems due to the use of incorrect criteria for stratification or the wish to use a different form of stratification to analyze the data after they have been collected.

Data Quality Objectives

- There are seven steps to the DQO process:
- 1. *State the problem*: Describe the problem, review prior work, and understand the important factors
- 2. *Identify the goals of the study*: Find what questions need to be answered and the actions that might be taken, depending on the answers

DQO

- 3. *Identify inputs to the decision*: Determine the data needed to answer the important questions
- 4. *Define the study boundaries*: Specify the time periods and spatial areas to which decisions will apply; determine when and where to gather data
- 5. *Develop the analytical approach*: Define the parameter of interest, specify action limits, and integrate the previous DQO outputs into a single statement that describes the logical basis for choosing among possible alternative actions
- 6. *Specify performance or acceptance criteria*: Specify tolerable decision error probabilities (probabilities of making the wrong decisions) based on the consequences of incorrect decisions
- 7. *Develop the plan for obtaining data*: Consider alternative sampling designs, and choose the one that meets all the DQOs with the minimum use of resources

Impact Assessment

- The before–after-control-impact (BACI) study • design is often used to assess the impact of some event on variables that measure the state of the environment.
- The design involves repeated measurements over time being made at one or more control sites and one or more potentially impacted sites, both before and after the time of the event that may cause an impact.

Methods

- A simple method that is valid with some sets of data takes the differences between the observations at an impact site and a control site, and then tests for a significant change in the mean difference from before the time of the potential impact to after this time.
- This method can be applied using the differences between the mean for several impact sites and the mean for several control sites.

Example of Water use Profile Study

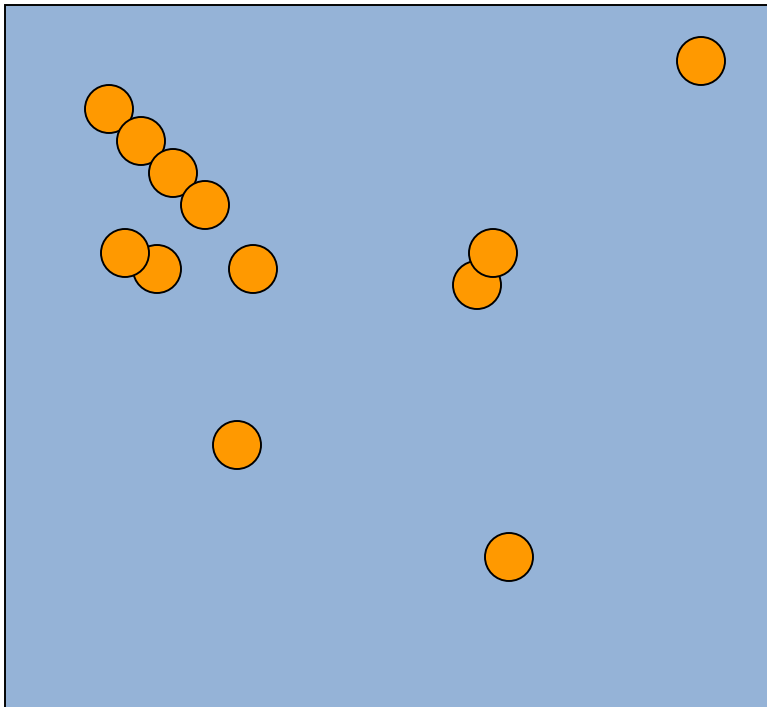
- Water Use profile study conducted in Broward County, Florida, USA.
- Need for evaluation of water usage among households in a region.
- Water meters need to be installed for assessment of water use.
- Need to develop a sampling strategy and assess the water use data.

Stratified Random Sampling

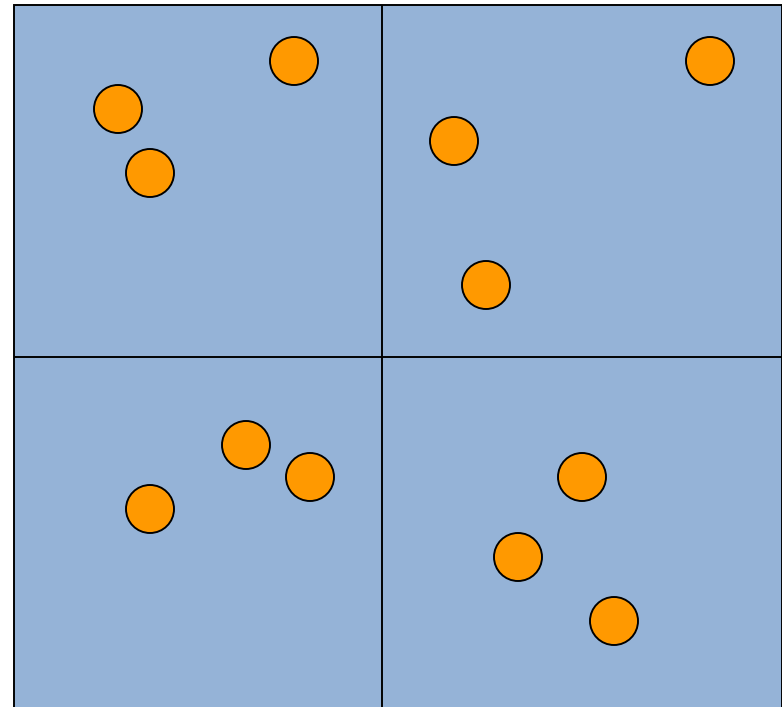
- The available population is first divided into a number of parts or 'strata' according to some characteristic (e.g. area, house type, zip code, etc in the current context), chosen to be related to the major variables being studied.
- Random samples are then selected from each stratum. The same proportion will be selected within each stratum, making the sample a proportionately stratified random sample

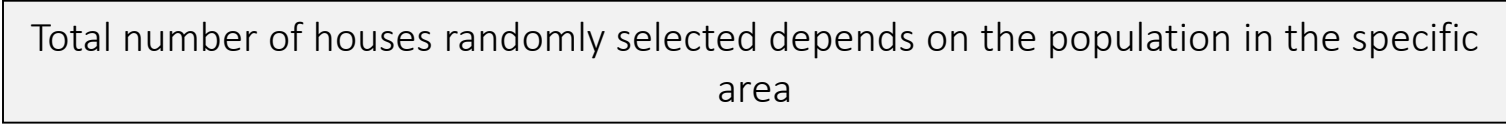
Sampling Schemes possible

- Random



- Stratified Random Sampling





Exercise

- In case of water profile study - think about how you can develop sampling strategy
- Would you go for a random sampling or stratified random sampling or systematic sampling ?
- If you decide to go one way or the other, how do you start selecting houses in a region ?
- What tool would you use to come up sampling sites (houses) ?
- How do you ensure that if stratified random sampling is selected you will obtain sites that will satisfy all the assumptions in your study ?

Sampling Schemes for Water Quality Assessment

- Sampling schemes that are adopted for identification of issues with environmental, hydrological and meteorological parameters are rarely random.
- Sampling sites are established primarily to assess constituents responsible or suspected to be causing problems, or affecting problem areas in planned in a well defined way.
- Sampling sites are established to understand the spatial variability of the constituent under investigation.

Sampling Sites

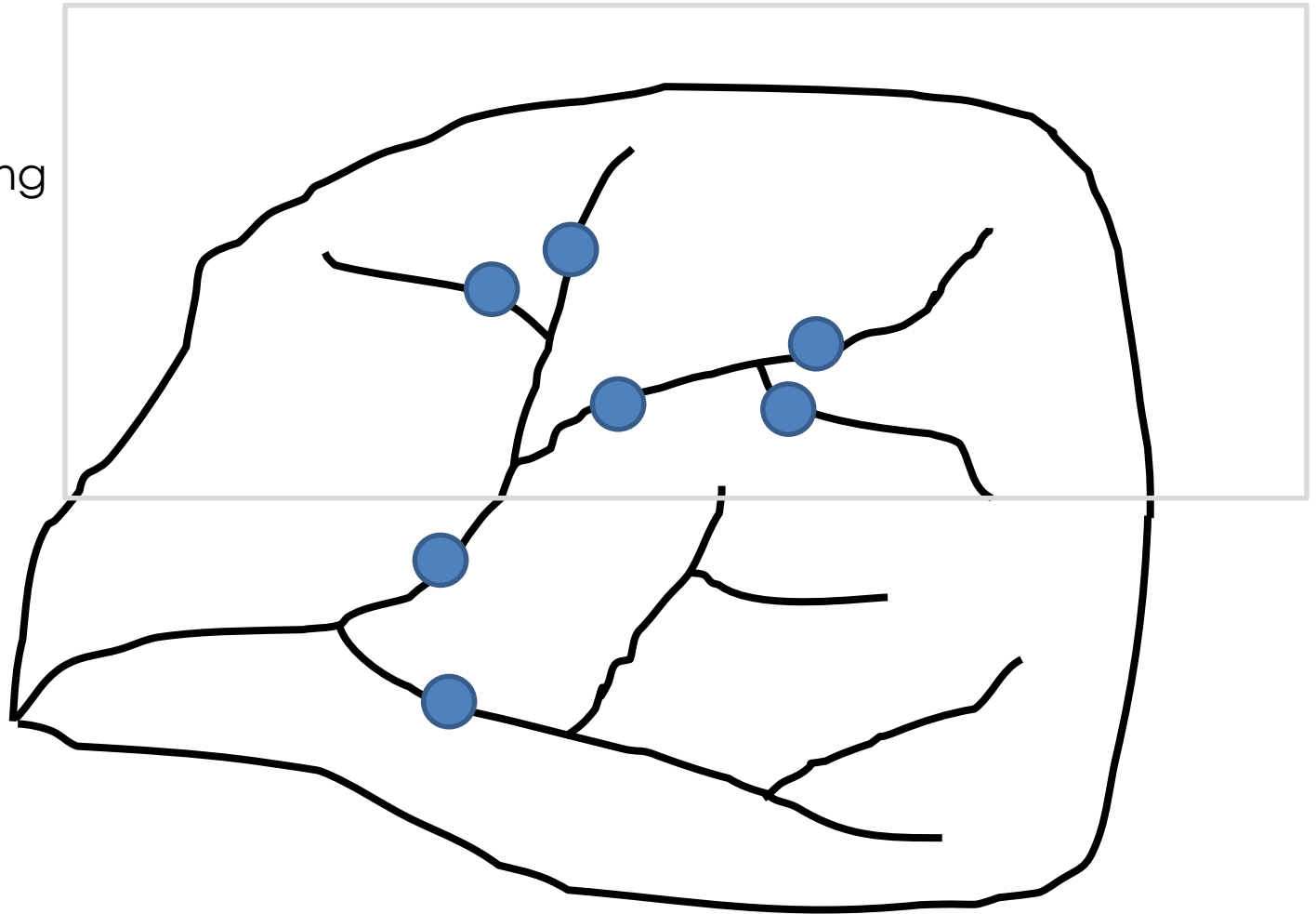
- Spatial Location is guided by
 - Problem areas
 - Objective-driven
 - Budgetary and other constraints
 - Logistics of sampling
- Examples (from Environmental field).
 - Sampling sites are established in areas where problems related to public health are reported.
 - Water samples are collected from rivers, streams, water storage areas if problems are suspected.

Sampling Sites

- Sites are decided not only based on the problem areas but:
 - Ease of sampling
 - Logistics
 - Ability to reach the spot and bring the sample to the laboratory for analysis within a reasonable amount of time
 - Ability to accurately sample in different times
 - Ability to measure all other parameters that are critical for evaluation of the collected samples
 - For example, water quality constituents often require discharge (e.g. flow rate) in the rivers when the sampling is conducted. This requires accurate measurement of discharge at sites. Sites at bridges crossing the rivers are often considered extremely beneficial for sampling purposes.

Example: Sampling Locations

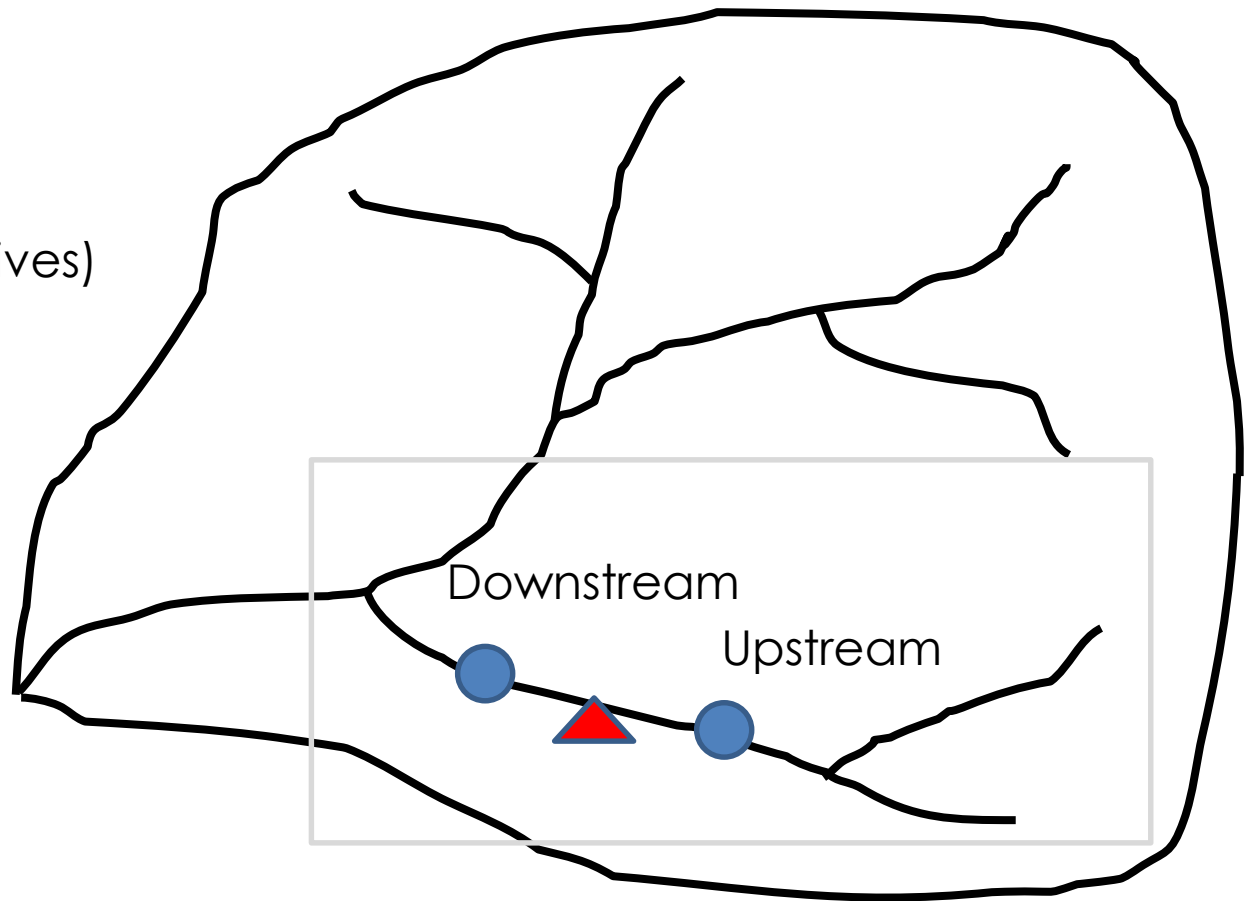
Exhaustive Sampling



Example: Sampling Locations

Project Specific

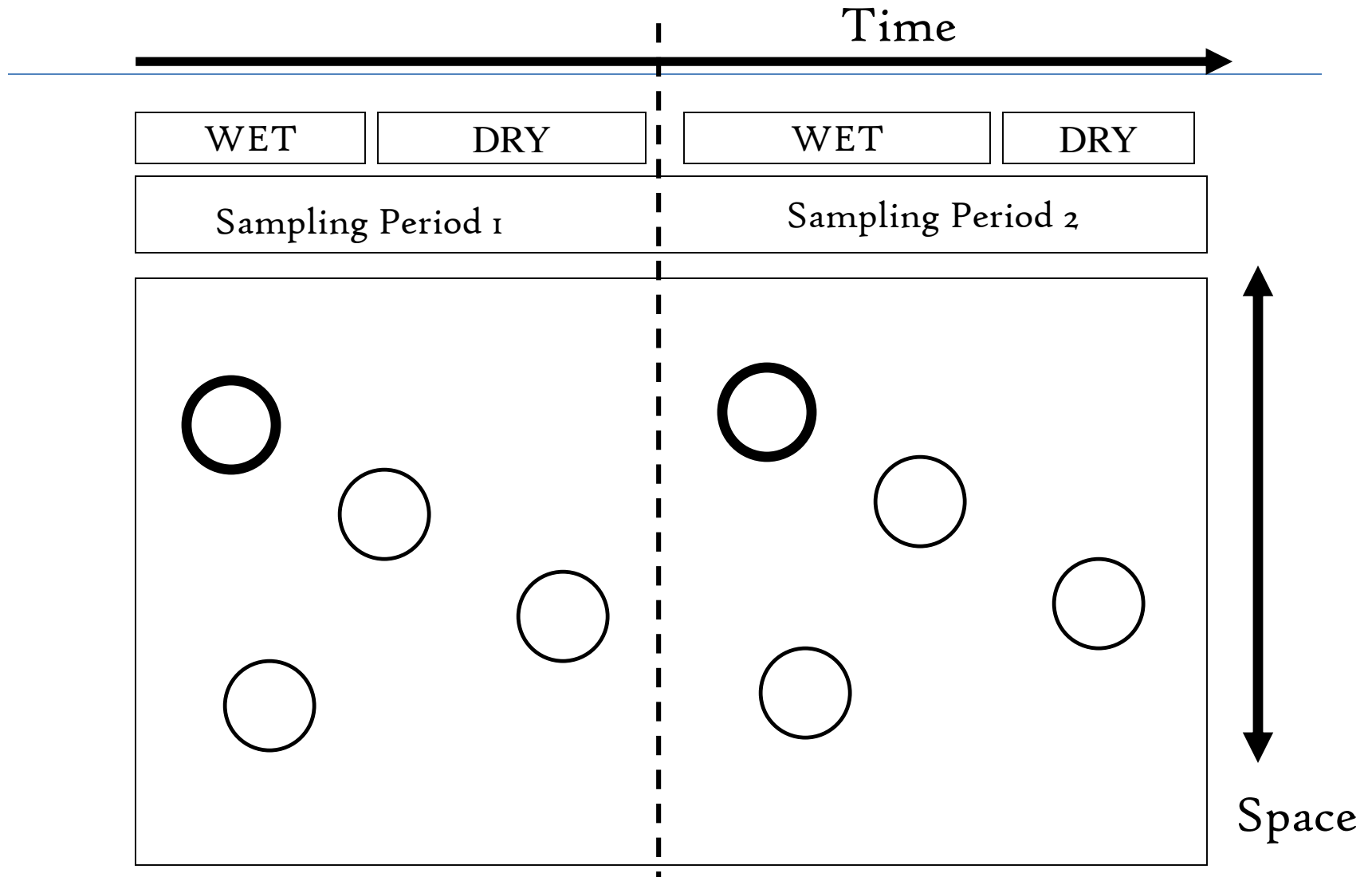
(Sampling with
Specific objectives)

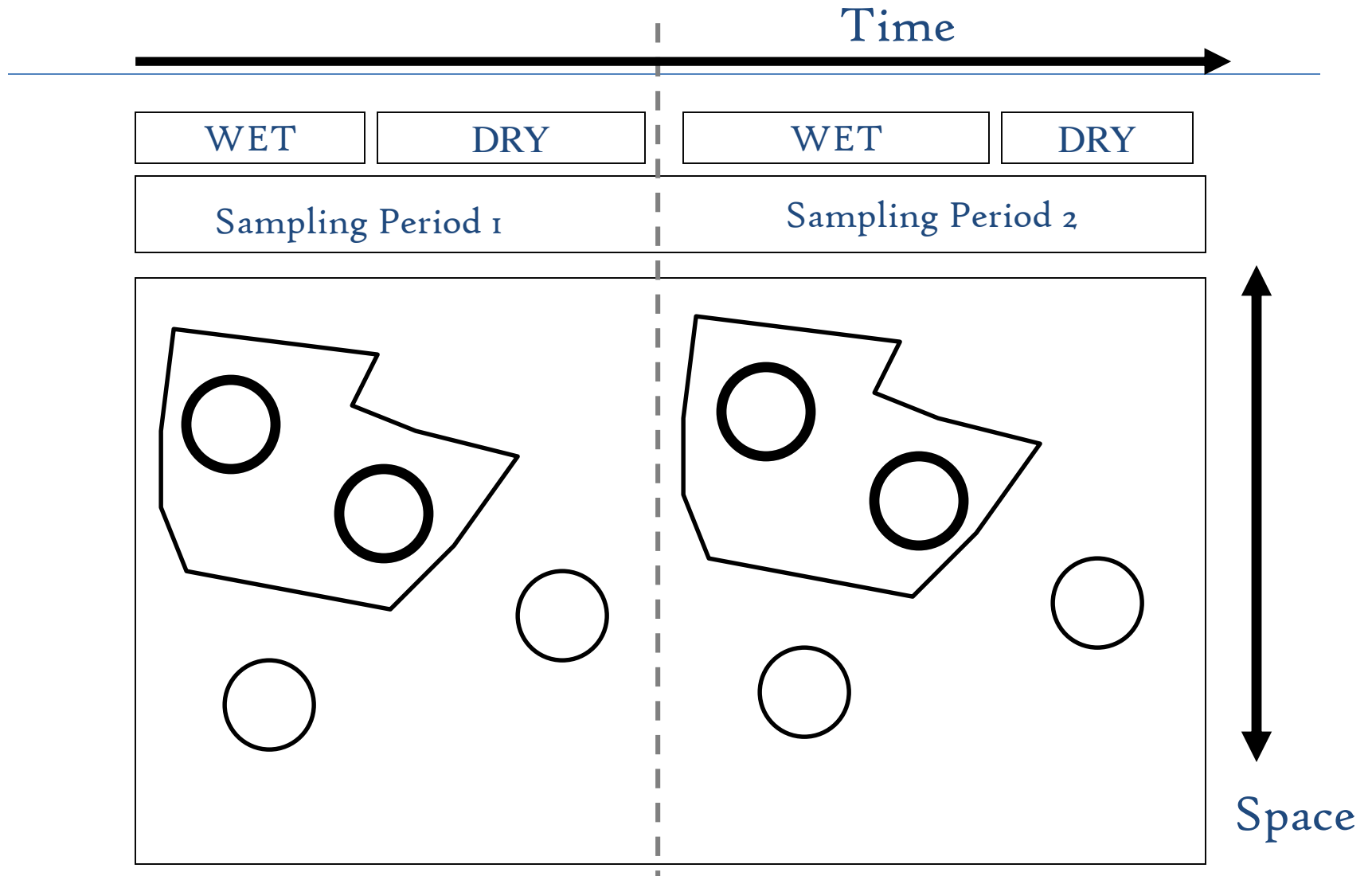


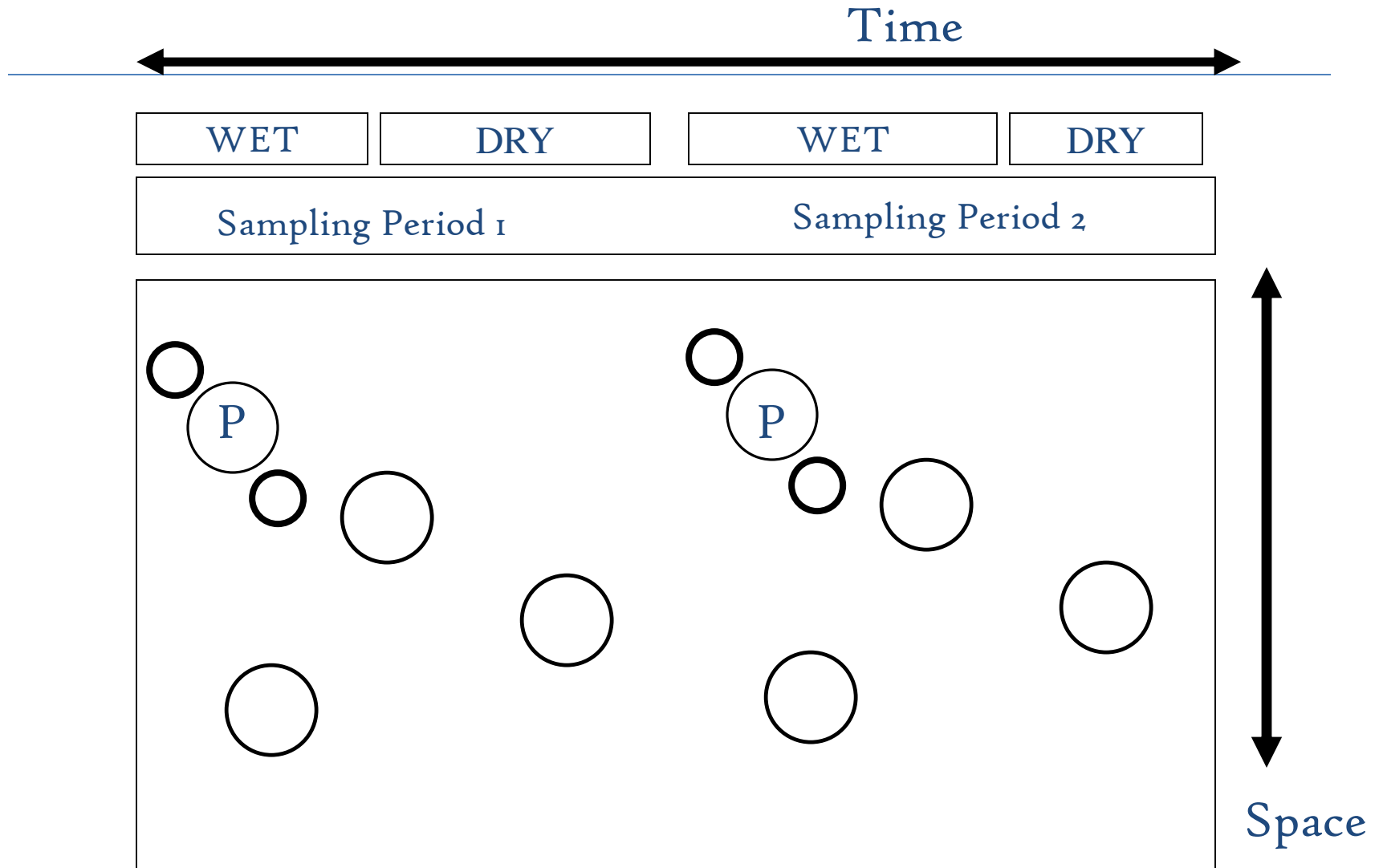
Sampling Strategies

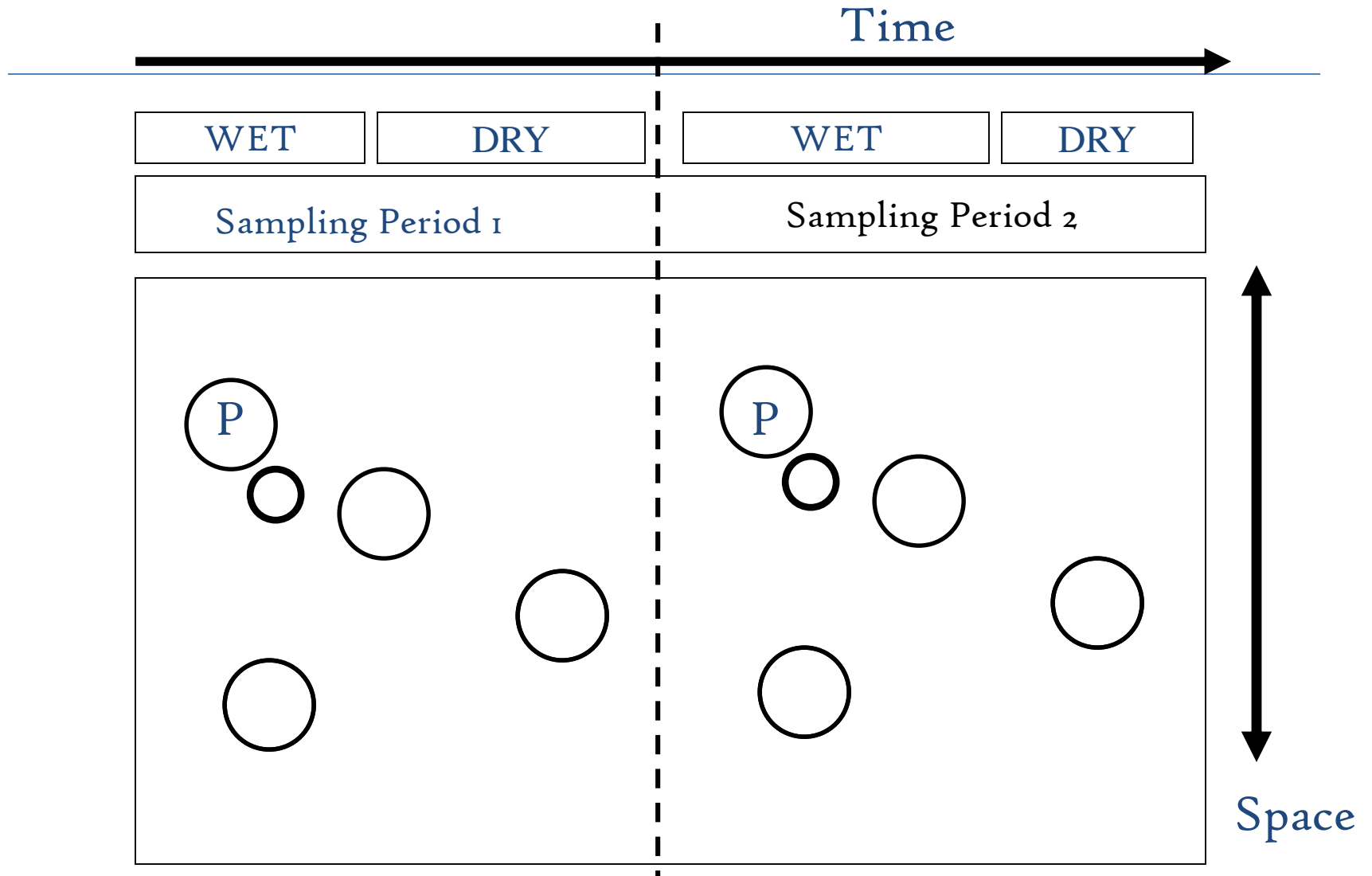
- Many time sampling site locations are determined by different agencies for different objectives. This results in clustering of the sampling locations.
- Clustering of stations results in redundant information if the phenomenon that is being investigated has a very low spatial variability – and does not warrant several stations at a short distance.

Examples of Water Quality Sampling site determination



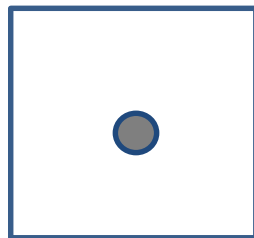






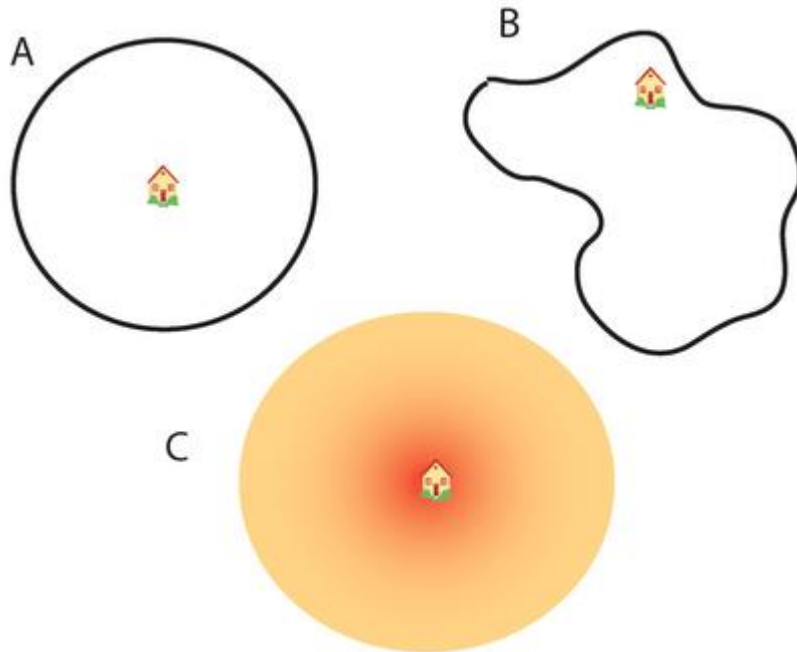
Spatial Relationships

- Co-location
 - When two objects of interest are located nearby (defined by a specific distance-based criterion) or point in polygon or sharing common area – then these two objects are known to be co-located.



Spatial Relationships

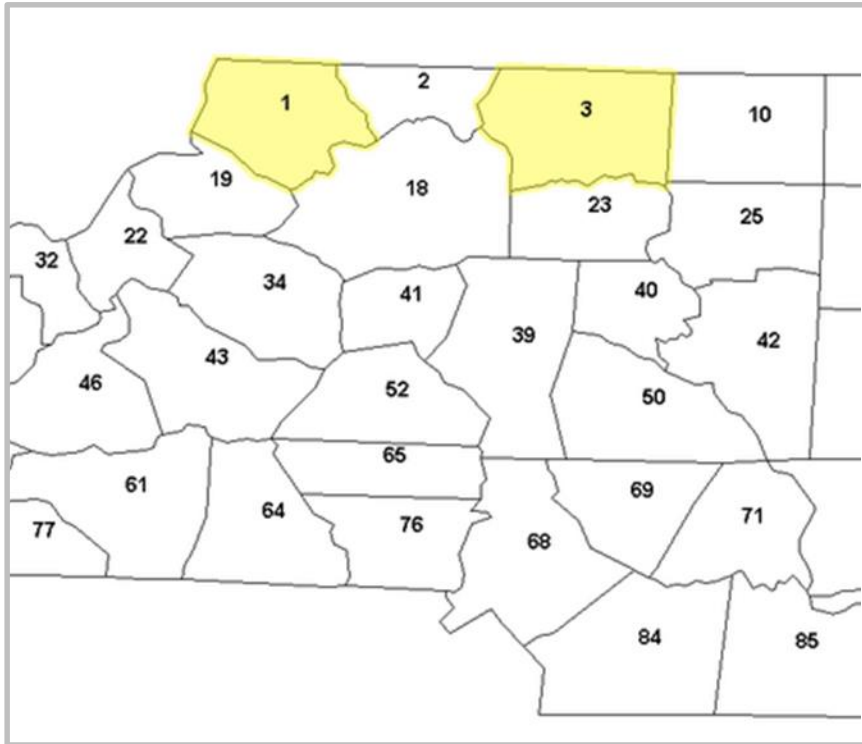
- Neighborhood



- Neighborhood of a house defined by :
- Fixed distance from the house (say a circle of specific radius)
- By a region, census area or sub-division of houses
- Defined by varying definition of nearness to different objects surrounding the house. Neighborhood defined by changing distance from the location of house.

-
- In the previous example : Neighborhood can be defined with weight decreasing as a simple function of distance.

Distance - Definition



- Many types of spatial analysis require the calculation of a table or matrix expressing the relative proximity of pairs of places, often denoted by **W** (a *spatial weights matrix*).
- Proximity can be a powerful explanatory factor in accounting for variation in a host of phenomena:
- Examples: flows of migrants, intensity of social interaction, or the speed of diffusion of an epidemic.

Spatial Data

- Pitfalls of Spatial Data
 - Spatial Autocorrelation
 - Positive spatial autocorrelation
 - Negative spatial autocorrelation
 - Non-correlation
- Modifiable Areal Unit Problem (MAUP)
 - **Ecological Fallacy**

Autocorrelation

- Autocorrelation **undermines** the conventional inferential statistics, due to redundancy in data arising from similarity in nearby observations.
- Tobler's law serves to help this issues

Scale

- Scale will have an important impact on spatial analysis, and choice of an appropriate scale is an important first step in all spatial analysis

Spatial Statistical Analysis

- Geographical data sets are not samples
- Geographical data are not random
- Because of autocorrelation, geographical data are not independent random (independent random process - IPR) or Complete Spatial randomness (CSR)
- Because n (sample size) is always large we will always find that the results are statistically significant
- What matters is scientific significance not statistical significance.
 - Peter Gould

Complete Spatial Randomness

- Example – Create a 2 –dimensional graph with X and y axis values ranging from 0 – 99.
- Pick two numbers randomly from a hat or a telephone directory between 0 and 99.
- What you have generated is a point pattern using a uniformly distributed random process.
- This can be regarded as CSR

Answers/Realizations

- The process-realization concept means that we can regard geographical data as samples in a very particular way.
- Geographical data **are not random.**
- Data not being independent does not prevent us from using statistics if we can develop better models than IPR/CSR

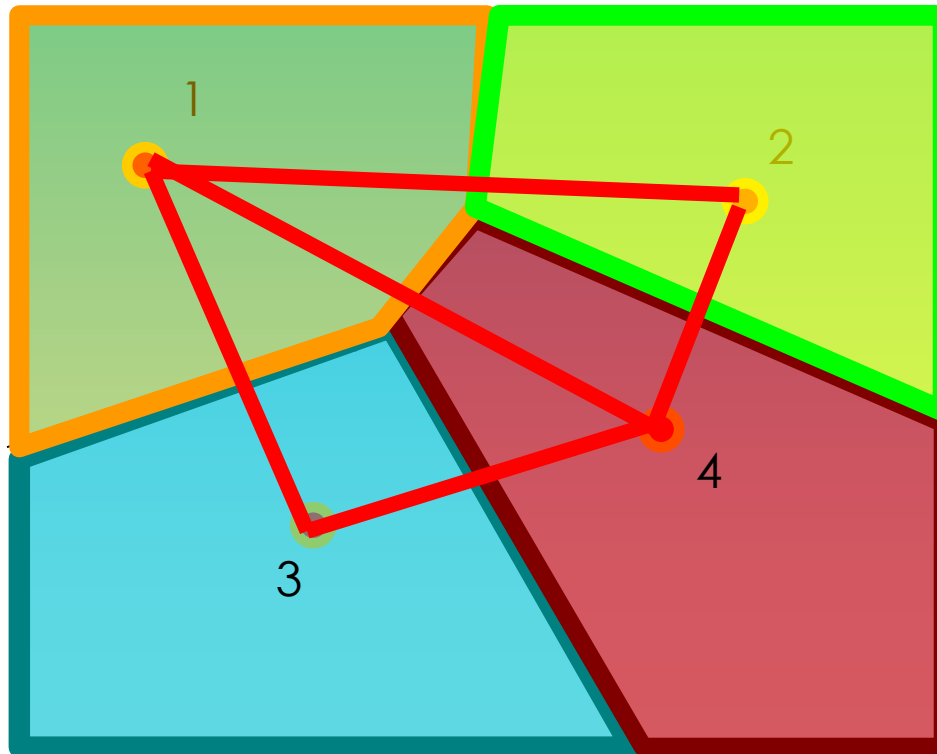
Proximity

- Proximity polygons and their duals the Delaunay triangulation are useful for construction in geographic analysis
- The variogram cloud and its summary (semi-variogram) provides a useful way to explore the spatial dependence between attributes and objects.

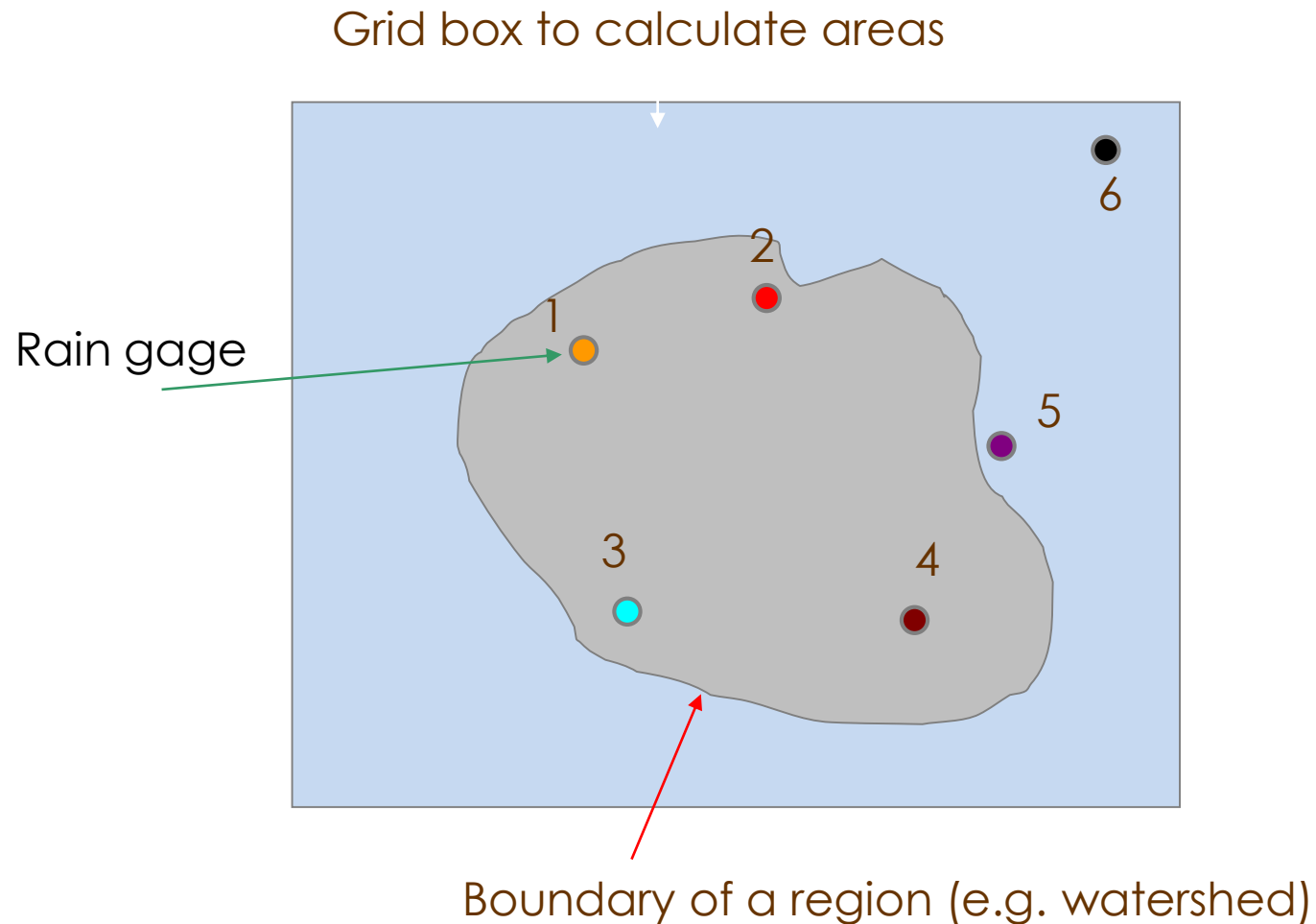
Thiessen **Polygons**

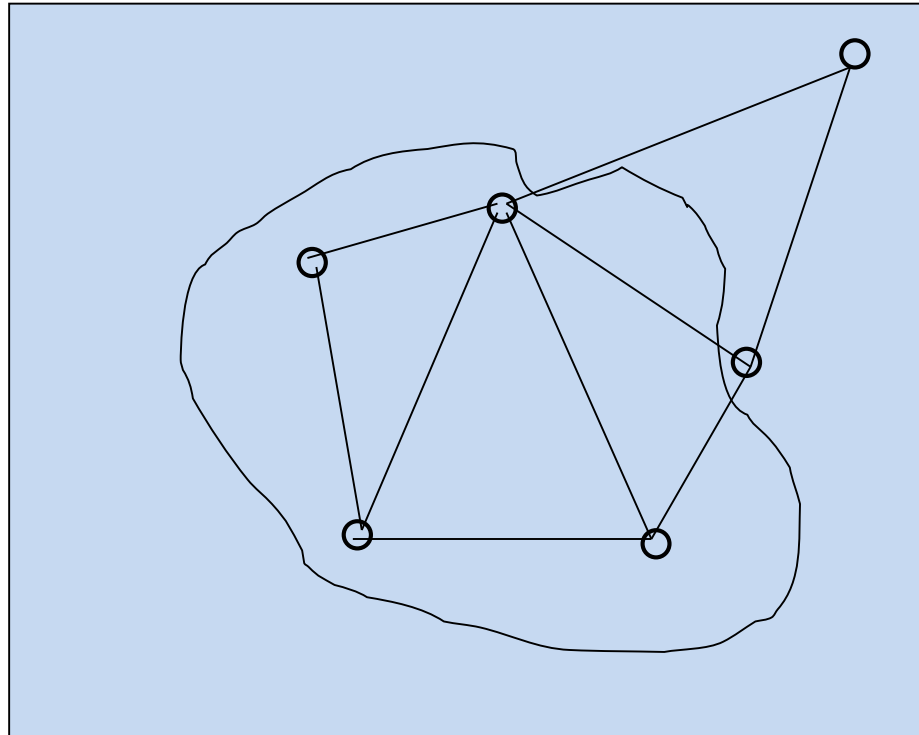
- **Polygons** generated from a set of sample points. Each Thiessen polygon defines an **area of influence** around its sample point, so that **any location inside the polygon is closer** to that point than any of the other sample points. Thiessen polygons are named for the American meteorologist Alfred H. Thiessen (1872-1931).

Proximity Polygons



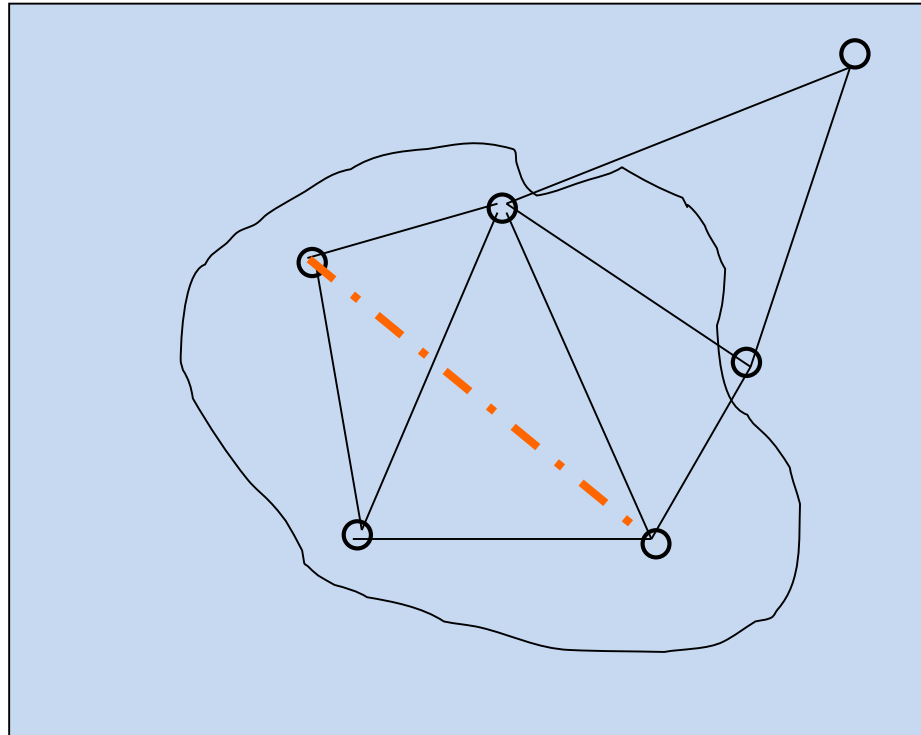
Application of Thiessen Polygon Concepts





Connect the ADJACENT rain gauges
Minimum number of triangles, use distances to decide

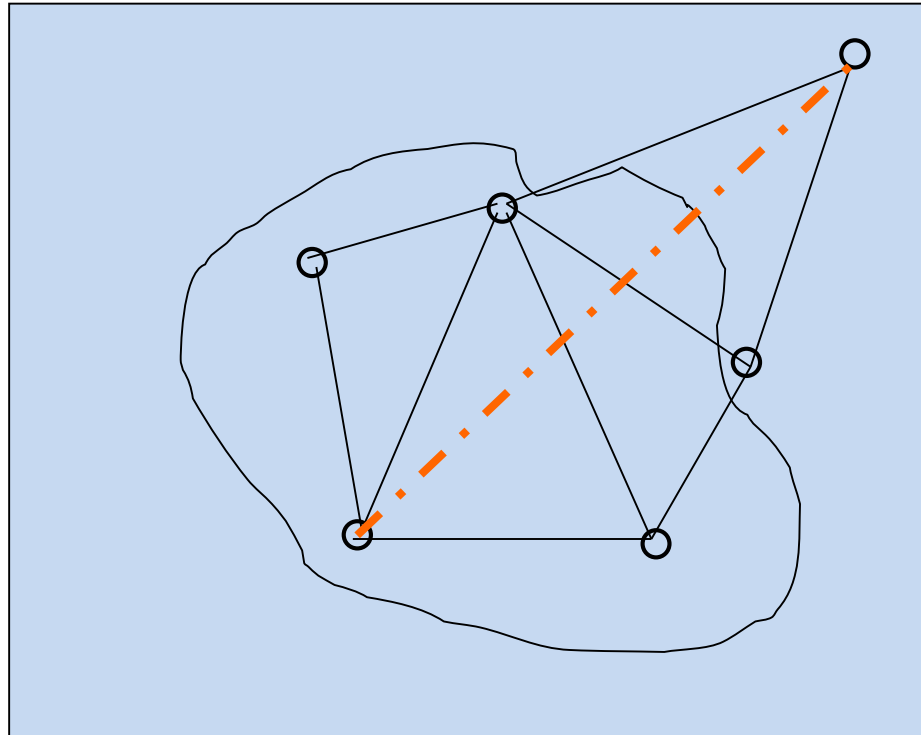
The orange
dotted
Line is crossing
the line
connecting
the other two
stations!!



NOT

ALLOWED

The orange dotted
Line is crossing
the line
connecting
the other two
stations!!



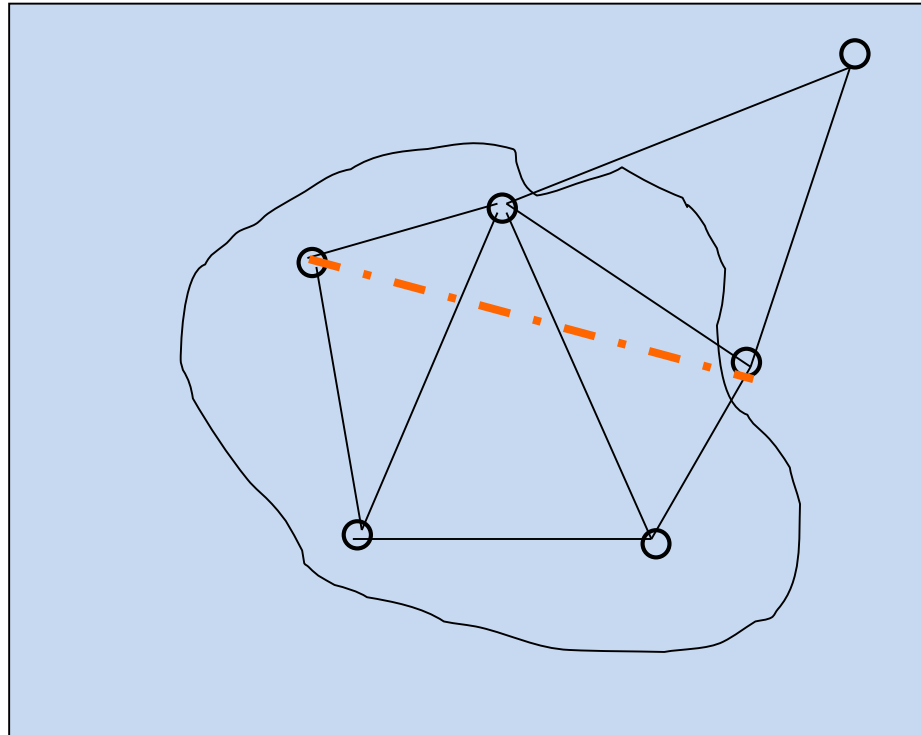
NOT

ALLOWED



The orange dotted

Line is crossing the line connecting the other two stations!!

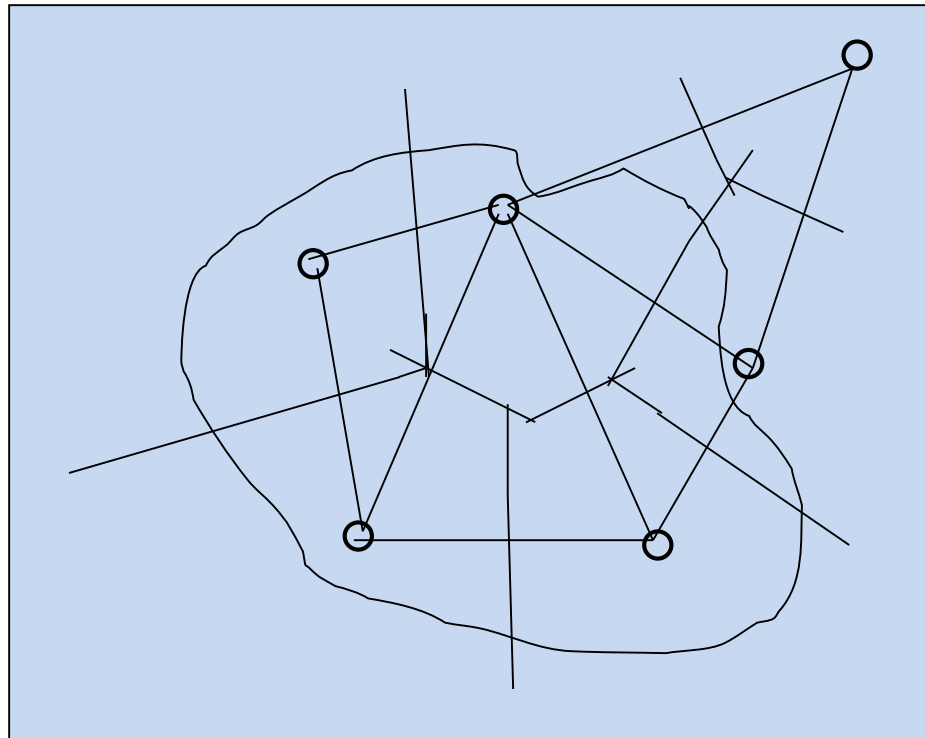


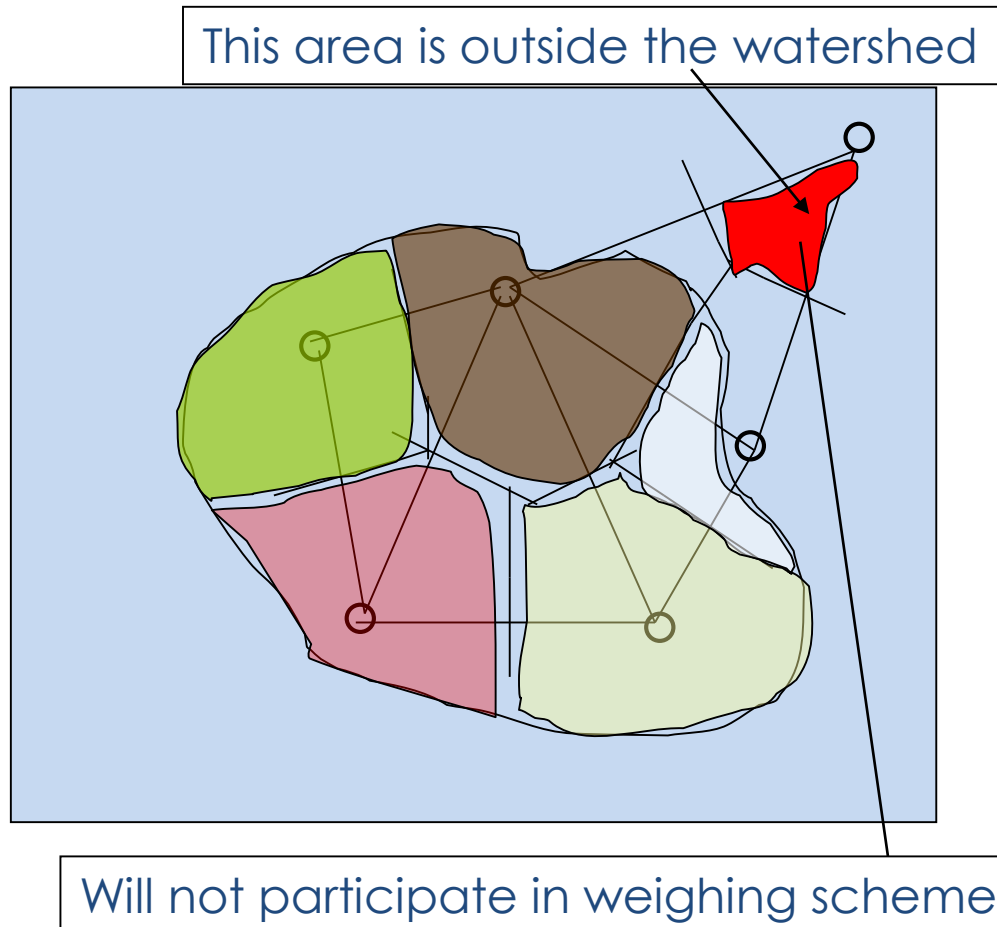
NOT

ALLOWED

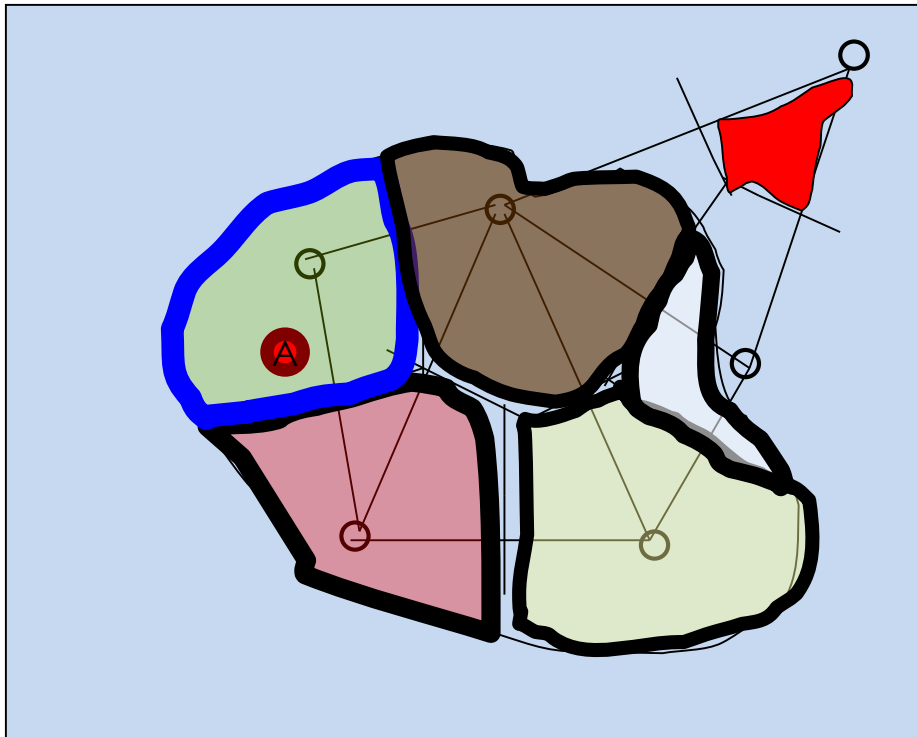
Draw Perpendicular Bisectors

Allow the perpendicular bisectors cross the watershed boundary





Property of Thiessen Polygon



ANY POINT
(e.g. point A)
IN a THIESSEN
POLYGON
(thick dark blue
boundary) is
closer to
STATION
(yellow colored
dot) than to
any other point
outside the
polygon

Common Rain Area Polygons

- It is safer to say based on the Thiessen polygon concept/property, any point in the polygon will have similar rainfall characteristics to those recorded at a rain gage located in that polygon.
- This property is used to find the average rainfall on a watershed using area-weighted rainfall concept.
- Thiessen polygons are also referred to as Voronoi or Proximity polygons
- These polygons are used often in spatial analysis, site location in geography (e.g. location of a fire station)

Thiessen Polygon Estimate of Average rainfall

$$P_{\text{average}} = \frac{\sum_{j=1}^N P_j W_j}{\sum_{j=1}^N W_j}$$

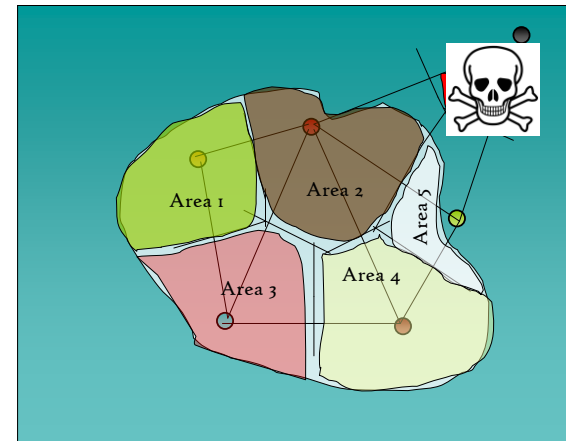
P_{average} = Average precipitation

P_j = Precipitation in area j

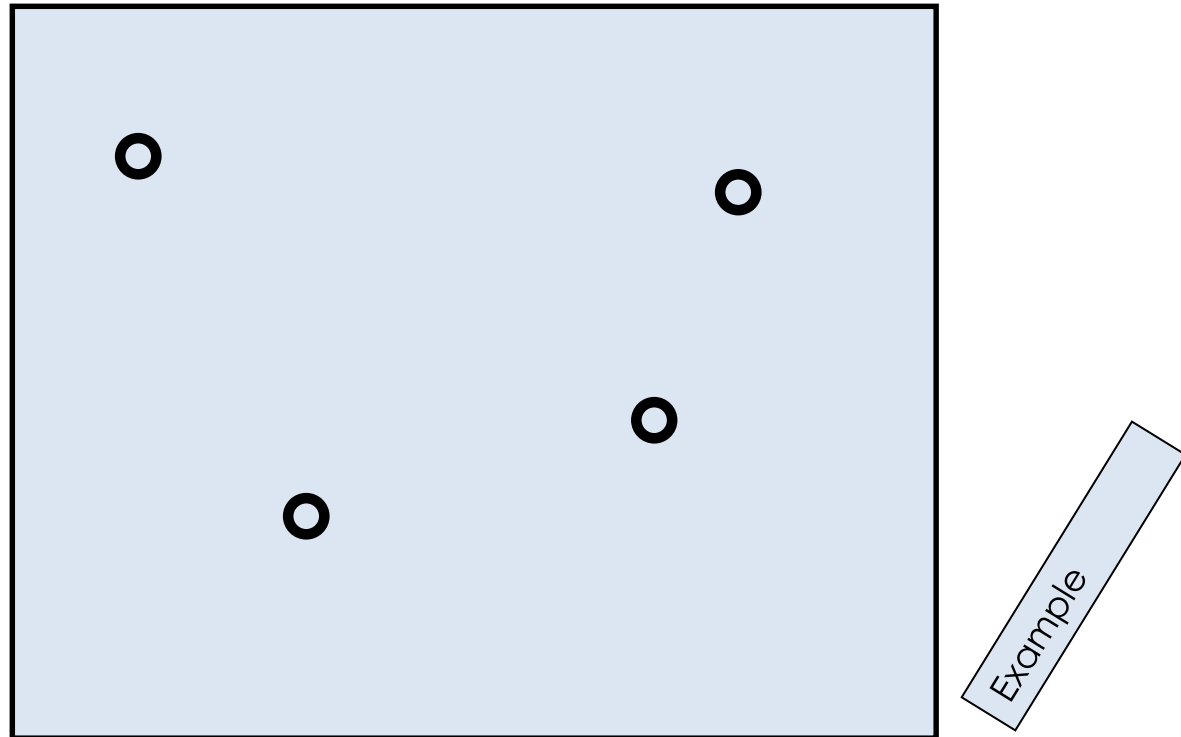
W_j = Area j

N = # of Areas

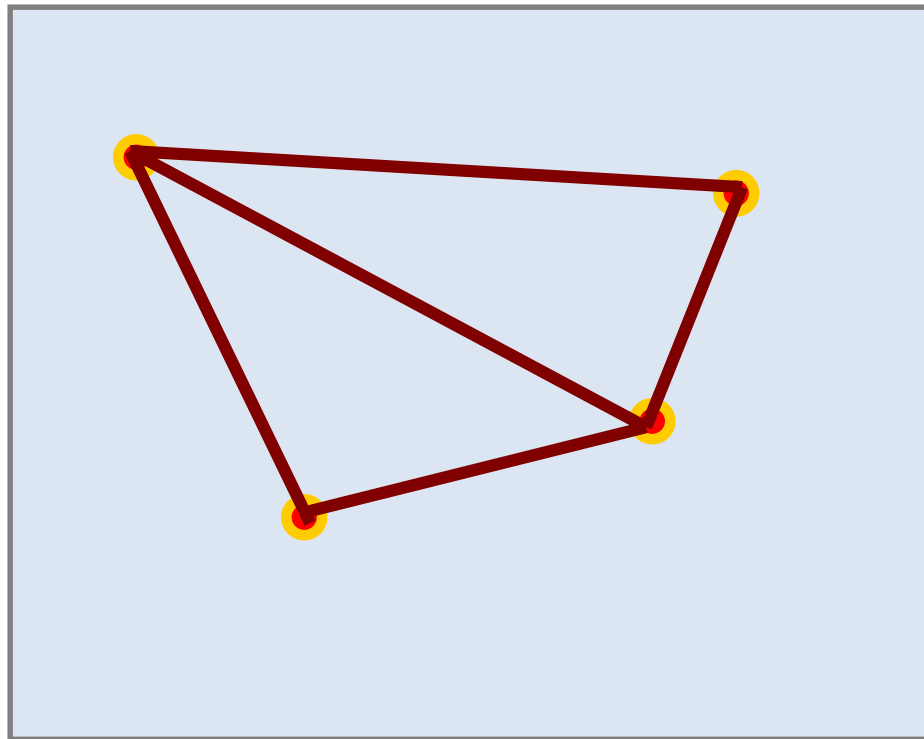
Area-weighted or Thiessen polygon method



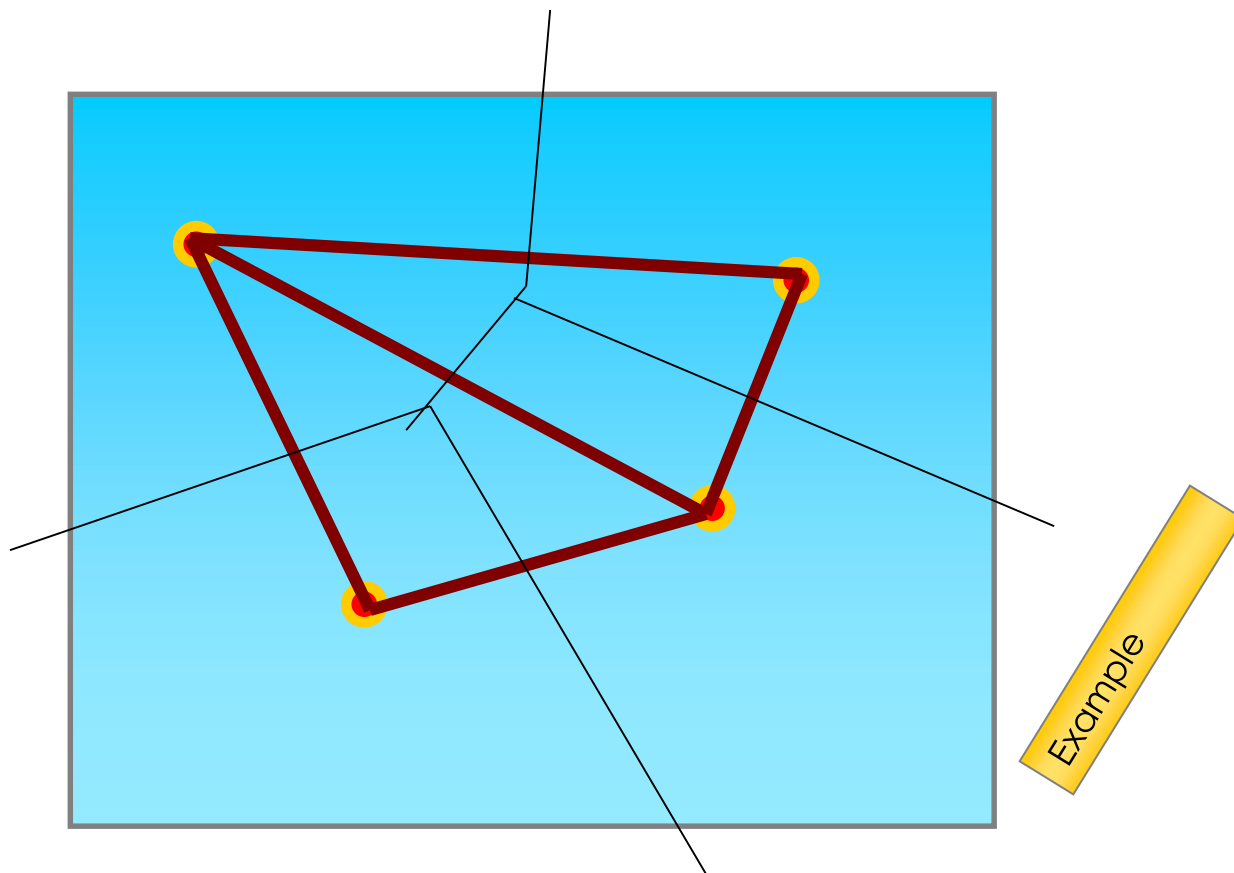
Sites in an area bounded by a rectangle



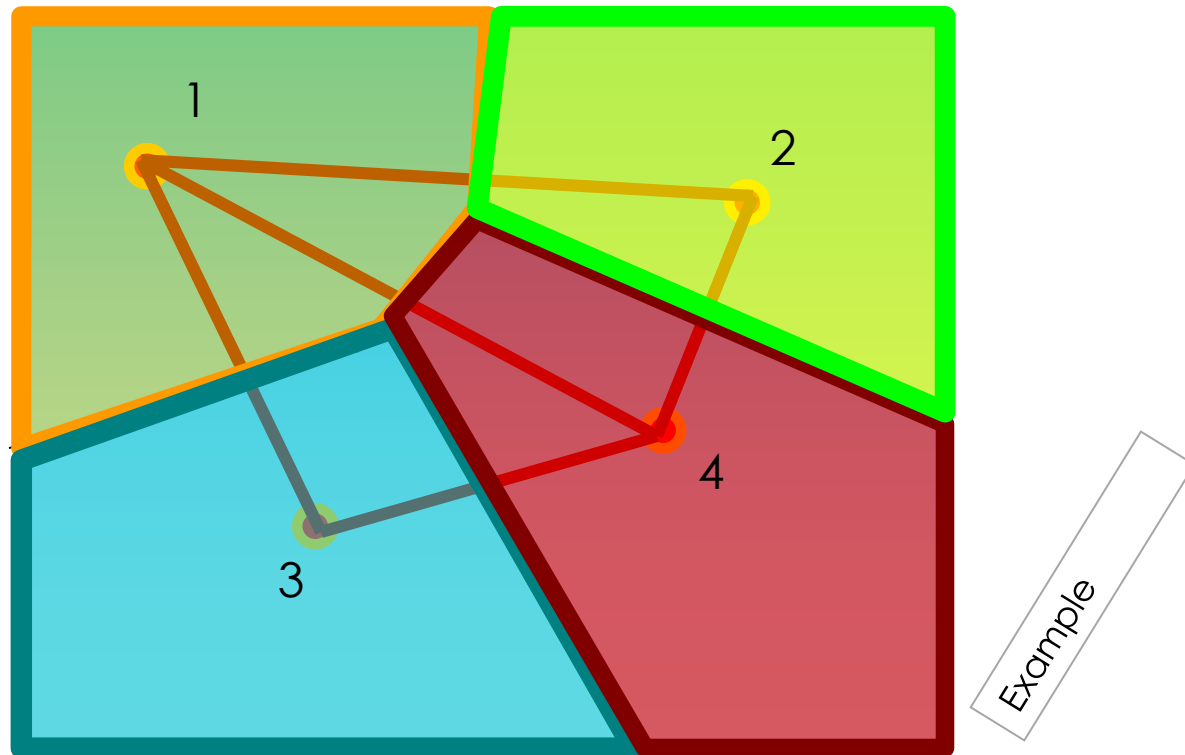
Triangles drawn



Perpendicular Bisectors Drawn



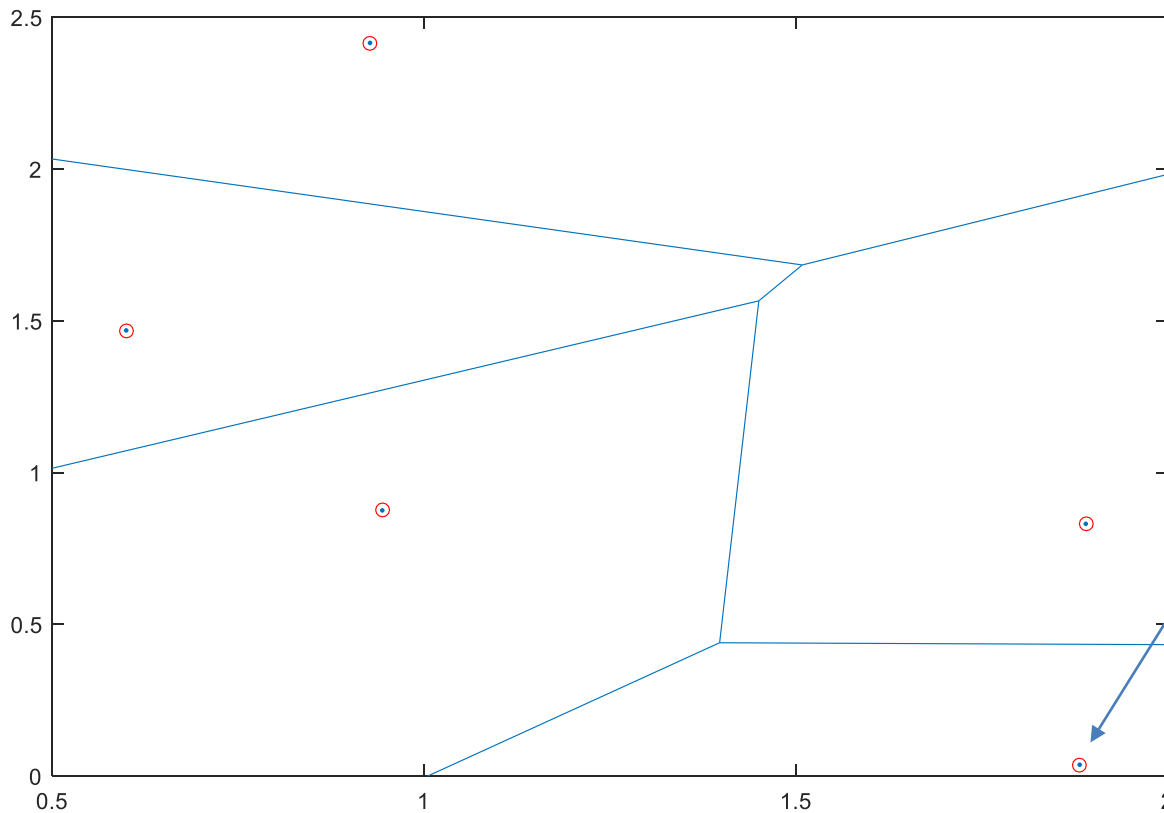
Thiessen Polygons



Thiessen Polygons. Exercises

- Developing Thiessen Polygons
 - Using MATLAB
 - ArcGIS
 - Or any other software

Voronoi Polygons



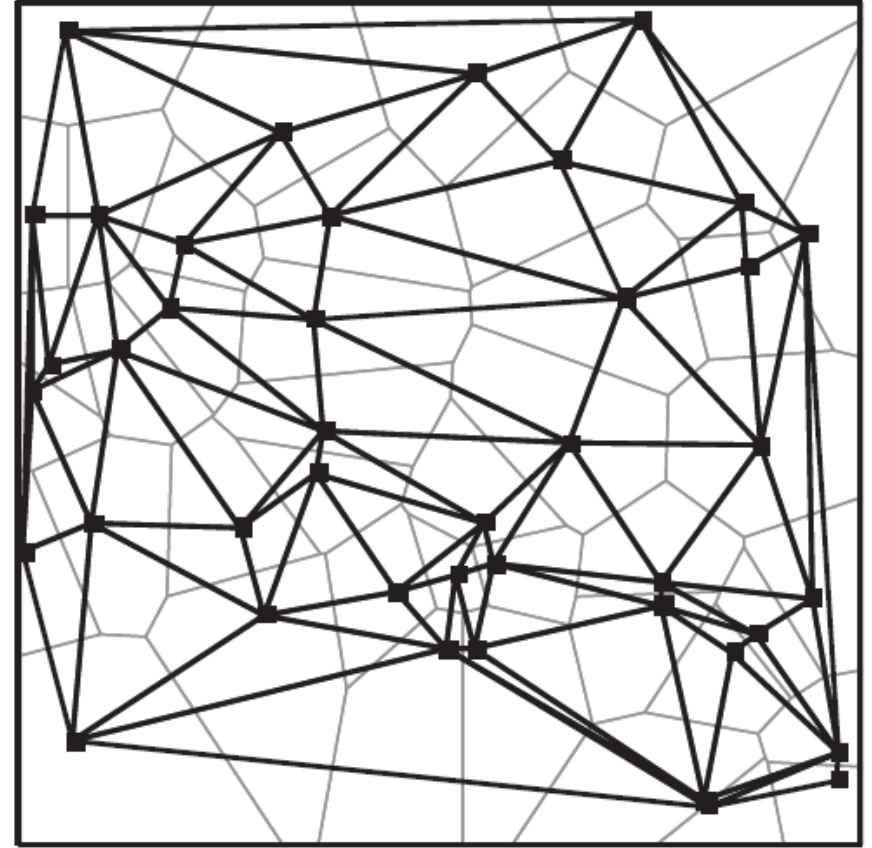
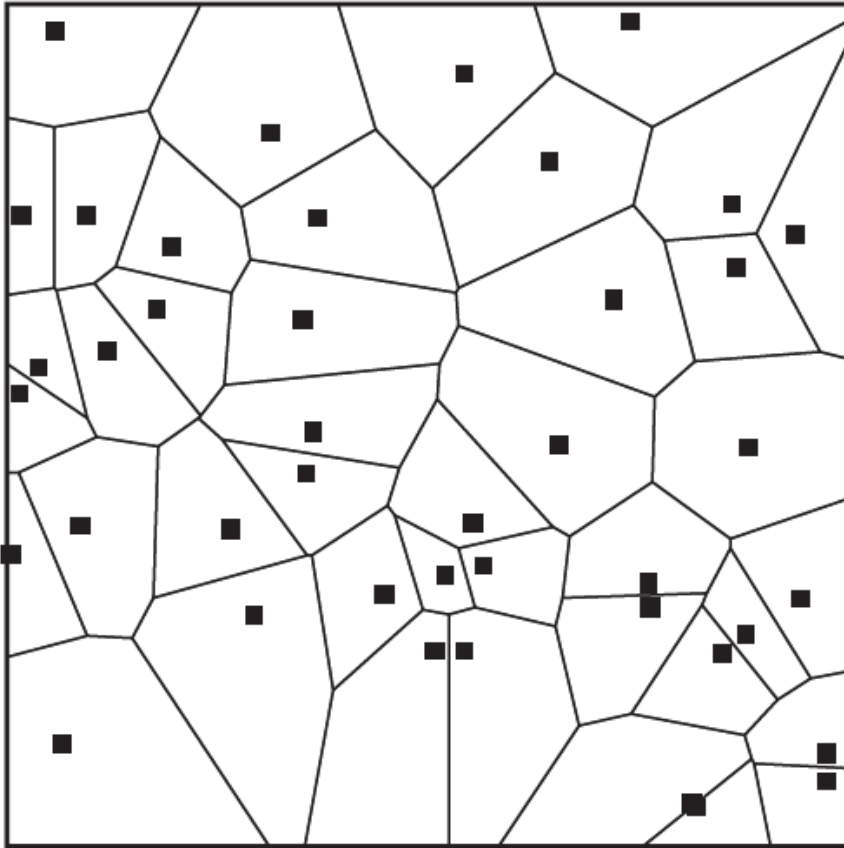
1.8818	0.0375
0.9440	0.8756
0.5997	1.4686
1.8899	0.8313
0.9275	2.4158

Using MATLAB

Proximity Polygons

- The proximity polygon of any entity is that region of the space which is closer to the entity than it is to any other.

Delaunay Triangulations



Delaunay Triangulation

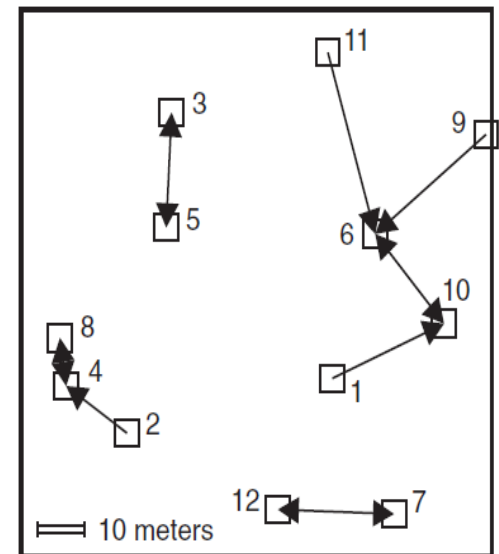
- Delaunay triangulation is derived from the proximity polygon by joining the pairs of events whose proximity polygons share a common edge.
- The distribution of areas in the proximity polygons are indicative of how evenly spaced (or not) the events.
- If the polygons are of similar sizes, then the events are evenly spaced.
- If the polygons are of not similar sizes, then the polygons of small sizes are likely to be in closely packed clusters.

Point Pattern Analysis

- Point density
- Point Separation
- Point density - First order Effect
 - Or intensity
 - Absolute location is very important
- Point Separation – Second order effect
 - Interactions between the locations
 - Relative distance is important

Point Pattern Analysis

Point	X	Y	NN	Distance
1	66.22	32.54	10	25.59
2	22.52	22.39	4	15.64
3	31.01	81.21	5	21.11
4	9.47	31.02	8	9
5	30.78	60.1	3	21.14
6	75.21	58.93	10	21.94
7	79.26	7.68	12	24.81
8	8.23	39.93	4	9
9	98.73	77.17	6	29.76
10	89.78	42.53	6	21.94
11	65.19	92.08	6	34.63
12	54.46	8.48	7	24.81



$$d(\mathbf{s}_i, \mathbf{s}_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Mean Nearest Neighbor Distance

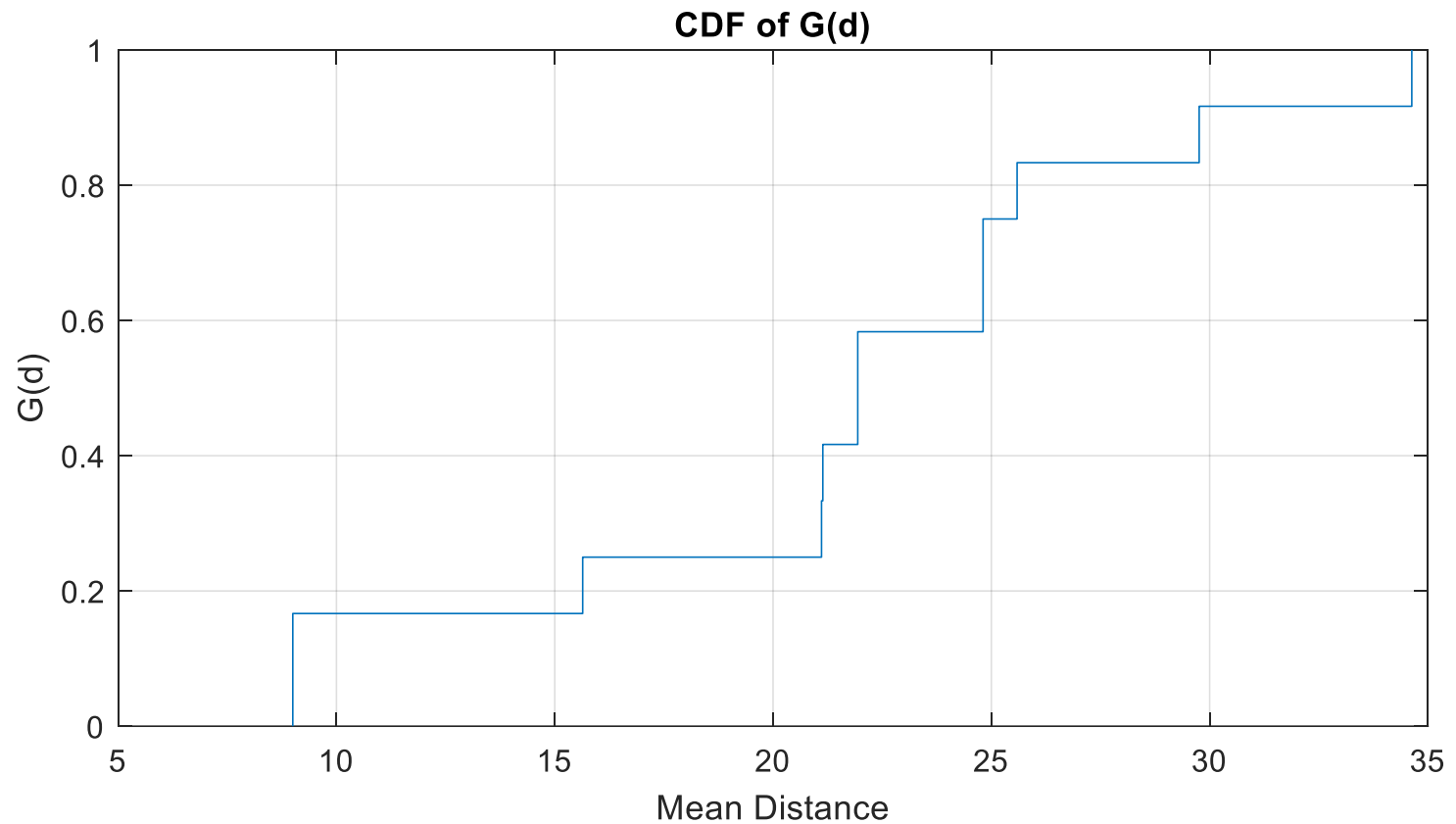
$$\bar{d}_{\min} = \frac{\sum_{i=1}^n d_{\min}(\mathbf{s}_i)}{n}$$

G- Function

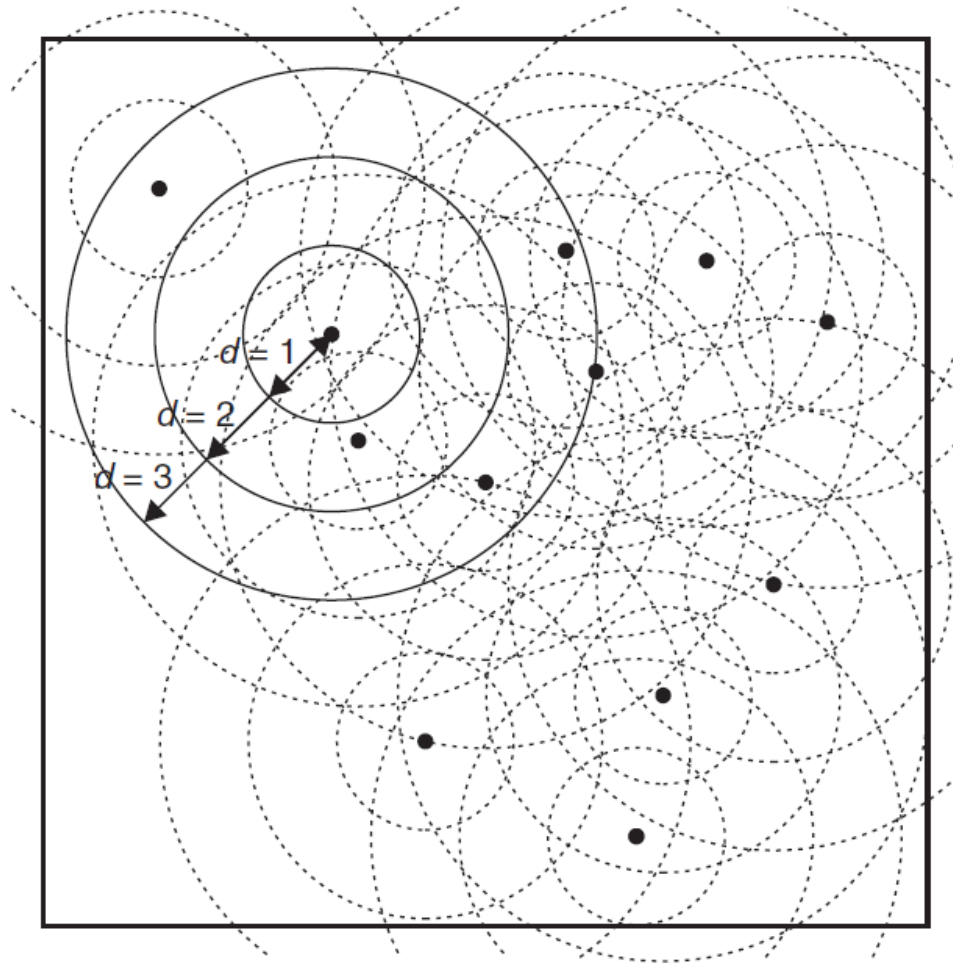
- G function, sometimes called the refined nearest neighbor.
- Instead of summarizing it using the mean, we examine the cumulative frequency distribution of the nearest-neighbor distances

$$G(d) = \frac{\#(d_{\min}(\mathbf{s}_i) < d)}{n}$$

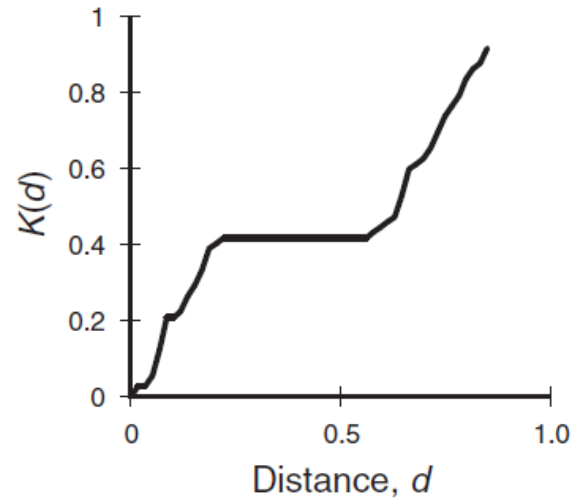
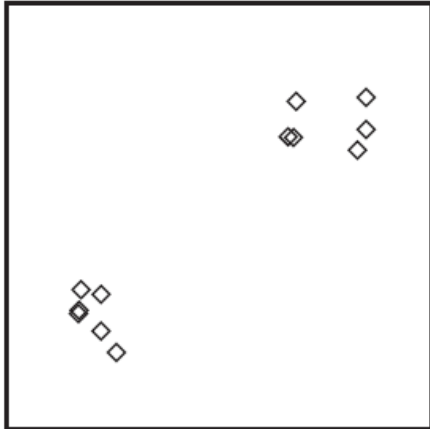
G-function



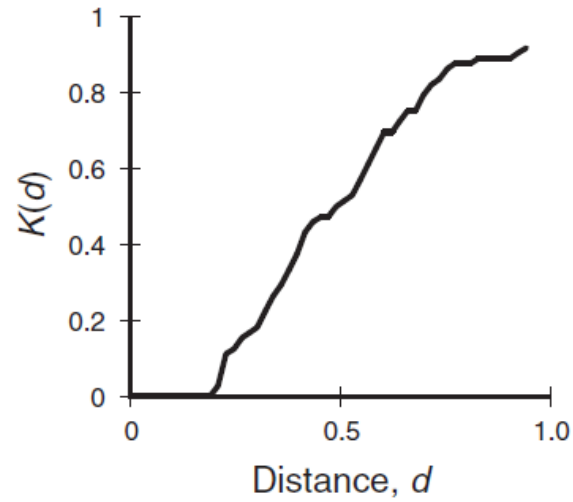
K-function Explanation



Clustered



Evenly spaced



IRP/CSR & Tobler's Law

- Tobler's first law of geography tells us that real-world geography **almost never conforms to IRP/CSR since** “Everything is related to everything else, but near things are more related than distant things.”

Examples

- First order
 - Clusters, example specific trees grow only in specific land/soil types
 - Diseases in some pockets of the city
- Second Order
 - Development of cities and trade-centers
 - Aggregation – Occurrence of one event prompts the other events.

Spatial Processes

- A spatial process is “first-order stationary” if there is no variation in its intensity over space.
- A spatial process is “second-order stationary” if there is no interaction between events
- The independent Random process (IRP) is both first order and second order stationary

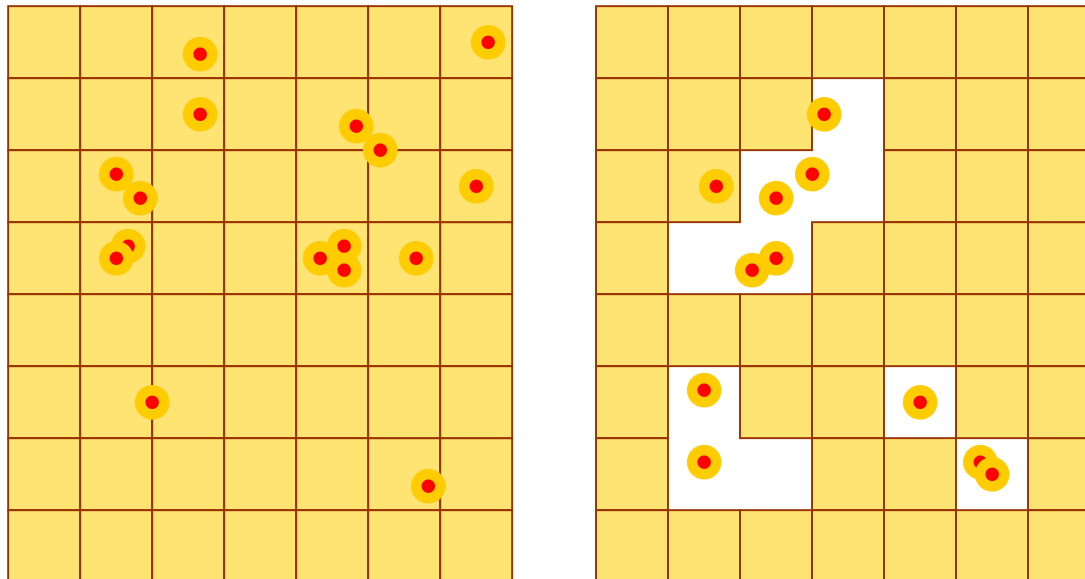
Intensity Variations

- Intensity variations can be occurring in one spatial direction only
- Spatial anisotropic.
- Isotropic processes are also possible in space.
- First order and second order effects are always part of the spatial processes that weaken the application of traditional spatial statistical methods.

Quadrat Counts

- Quadrat counts based on either census or a sample of quadrats provide a good, simple summary of point distributions
- These measures are subset of intensity or simple density measurements.
- Kernel density estimation assumes that the density is meaningful everywhere in a study region and is calculated by traversing the study region using a function that depends on a central location and the number of events around that location.

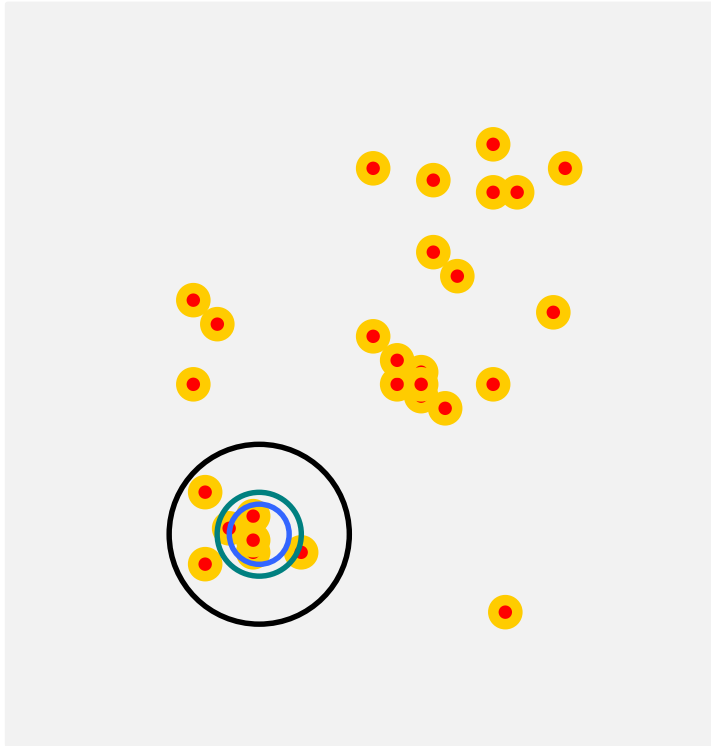
Quadrats and Census



Distance based point pattern measures

- Density based or intensity based approaches are based on the need for assessing the first-order processes
- Distance based measures are used to assess the second-order processes.
- Nearest-neighbor distance
 - (basically Euclidean Distance)
- G, F and K functional measures

Ripley K-function



- This function will help us to assess the randomness of a point distribution over different spatial scales.
- Assessment of K function involves several steps.
- Requires a definition of "distance" increment.

Steps for K function

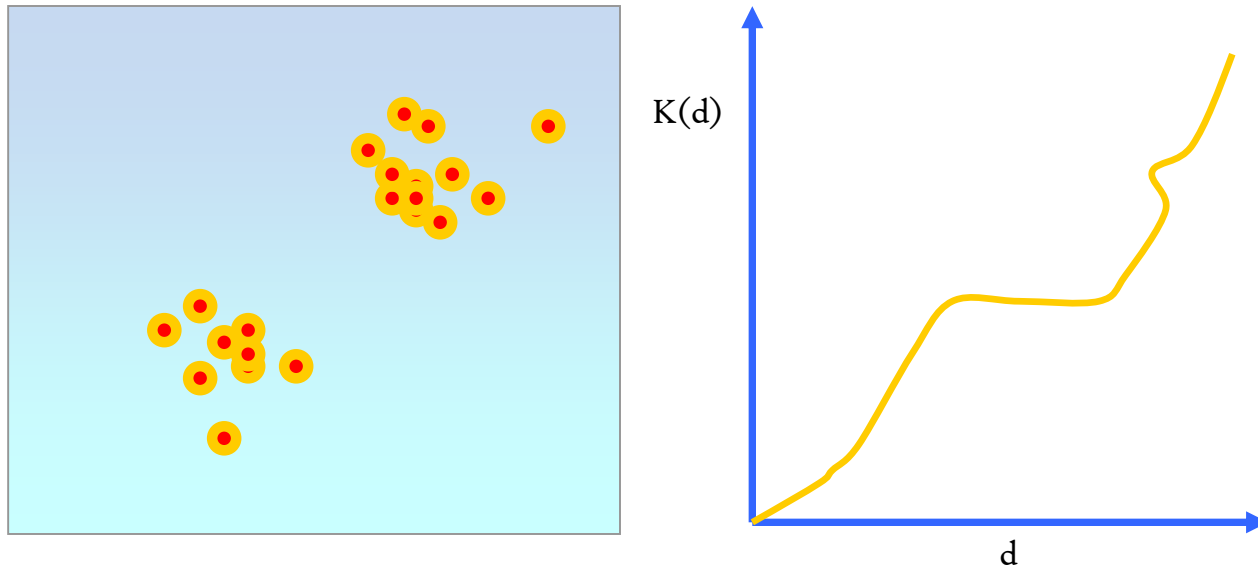
- Select a distance increment or spatial lag, “d” (similar to the units of the spatial scale)
- If the two farthest points in the region are “D” units apart then $d \ll D$.
- The number of lags = D/d
- Step one, $g=1$ (iteration variable)
- Step two
 - Around each point, i , in the region create a circular buffer with a radius of h , where $h = d * g$
 - The buffer will have a size d in the first iteration and $2d$ in the second and so on.

K-function Steps

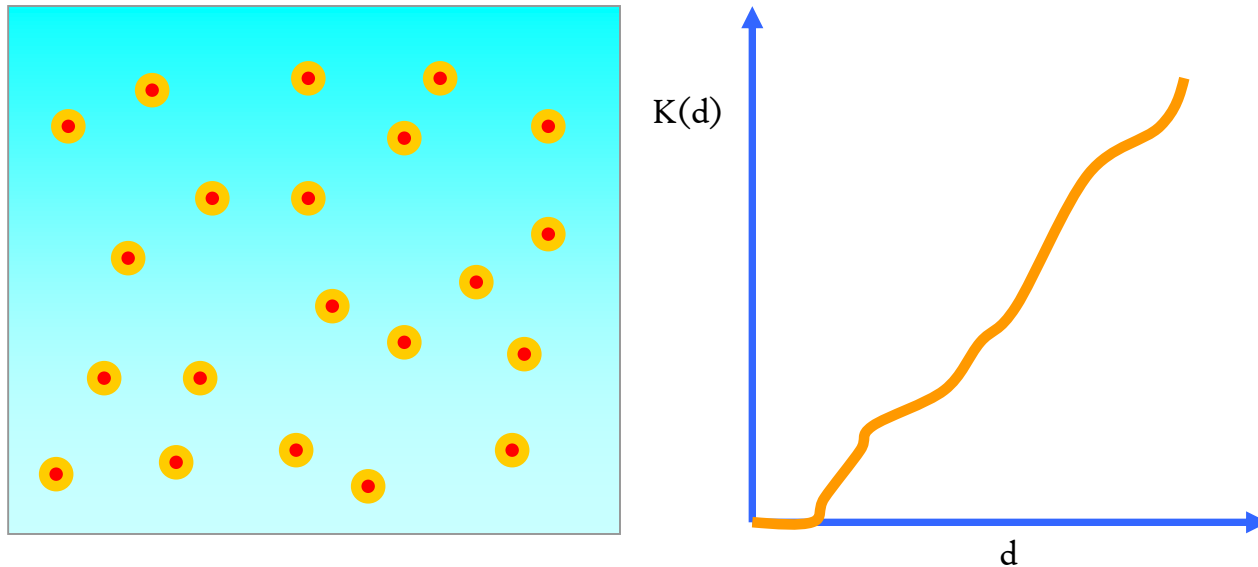
- For each point count the number of point falling within its buffer of size “h” and denote the count by $n(h)$
- Increase the radius of the buffer by d
- Repeat steps by increasing h until $g = r$ or $g = D/g$

-
- If the points form a highly clustered pattern, then high count values are within small “h” values and the counts will not increase much when the h is increased to relatively large sizes.
 - If the pattern is relatively dispersed the counts will be relatively low with small “h” values and increase quickly as “h” increases

K function for clustered events



K-function for evenly spaced events



Classification and Clustering

- Often times it is necessary to **classify** a set of samples collected or observed in space and time into a finite number of groups.
- **Clustering** also serves the purpose of establishing clusters (i.e. of groups of observations in space or time with similar properties)

Differences between Clustering and Classification

- Classification is a **supervised** process
- Clustering is an **unsupervised** process
- Classification is a process of categorization where objects are recognized, differentiated and understood on the basis of the training set of data.
- Classification is a supervised learning technique where a training set and correctly defined observations are available.

Differences between Clustering and Classification

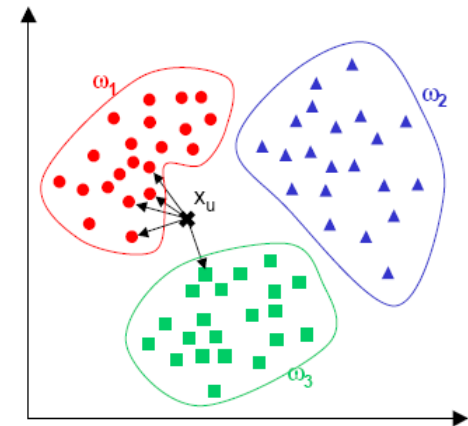
- Clustering is a method of grouping objects in such a way that objects with similar features come together, and objects with dissimilar features go apart. It is a common technique for statistical data analysis used in machine learning and data mining.
- Clustering can be used for exploratory data analysis and generalization.

K-Nearest Neighbor (KNN)

- The ***k*-nearest neighbor algorithm** is one of the simplest algorithms used for classification of spatial data into clusters
- An event/point or an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors.
- *The variable *k** is a positive integer, typically small. If $k = 1$, then the object is simply assigned to the class of its nearest neighbor.
- In binary (two class) classification problems, it is helpful to choose *k* to be an odd number as this avoids tied votes.

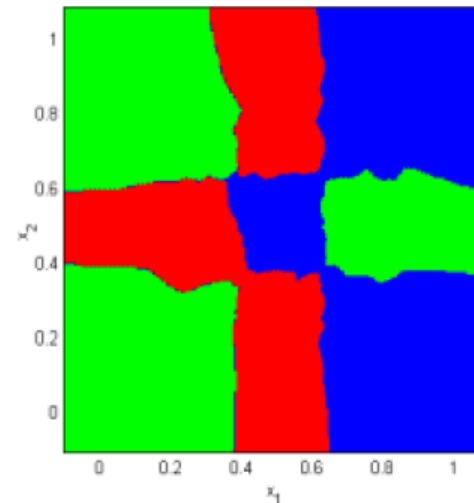
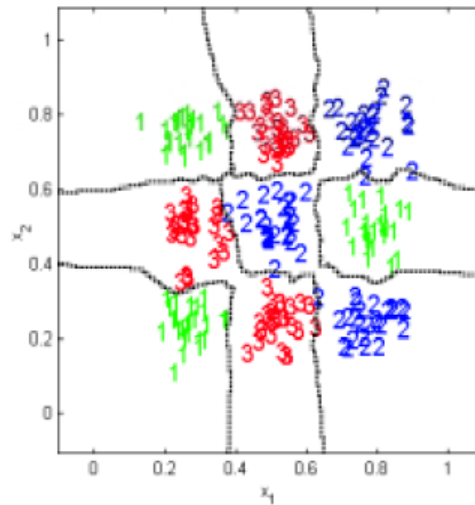
Classification

- Here $k = 5$,
- The 5 nearest neighbors from point are evaluated based on Euclidean distance.
- If majority of the nearest Neighbors are close to one Cluster, then that point Belongs to that cluster

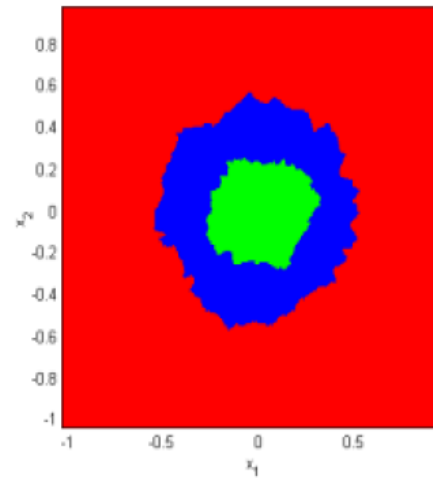
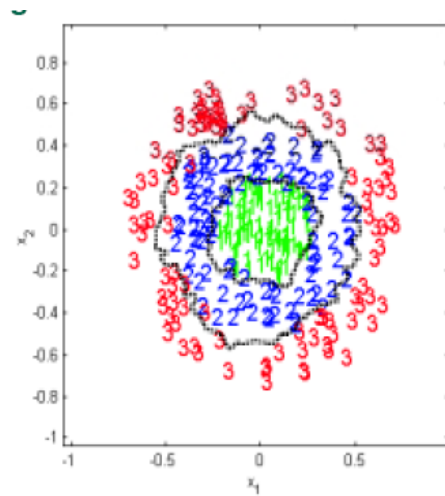


Ricardo Gutierrez-Osuna

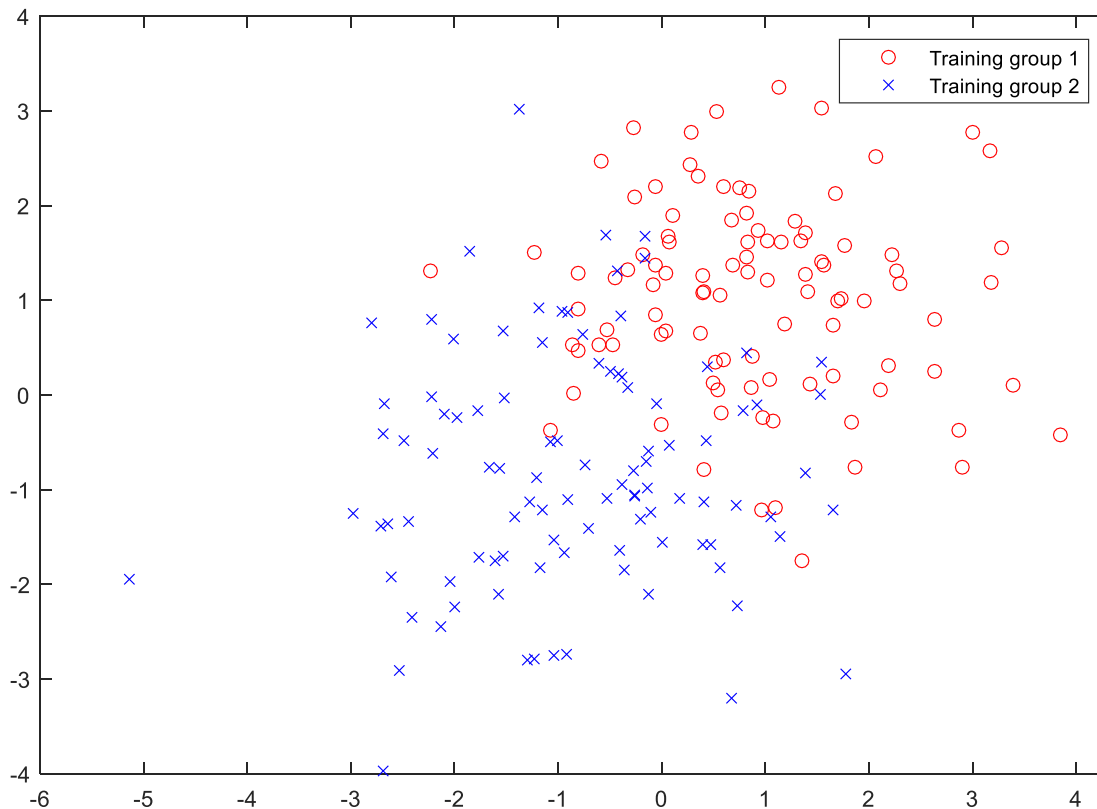
2- Dimensional classification problem (6 classes)



3 Classes

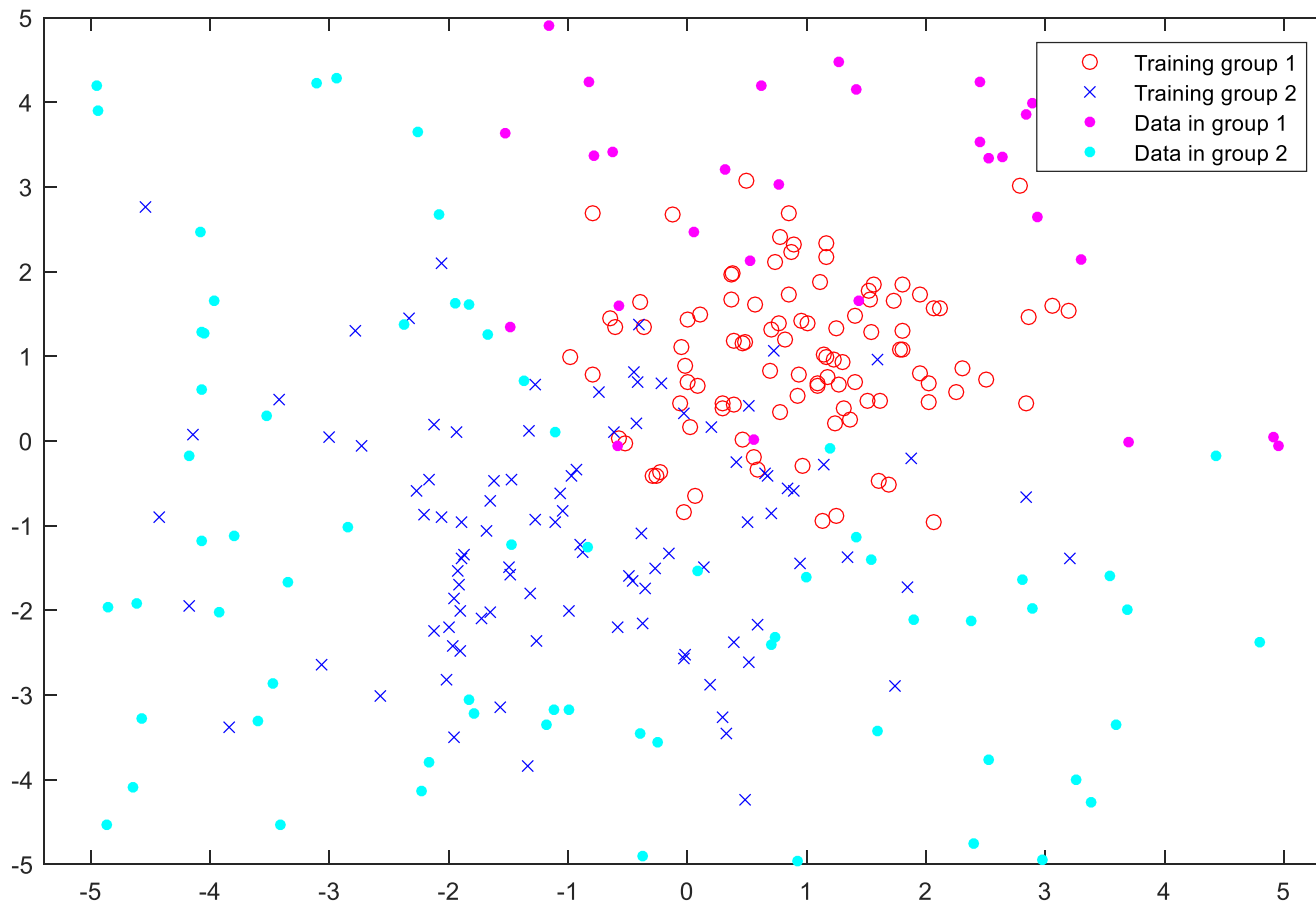


Example of two groups of data

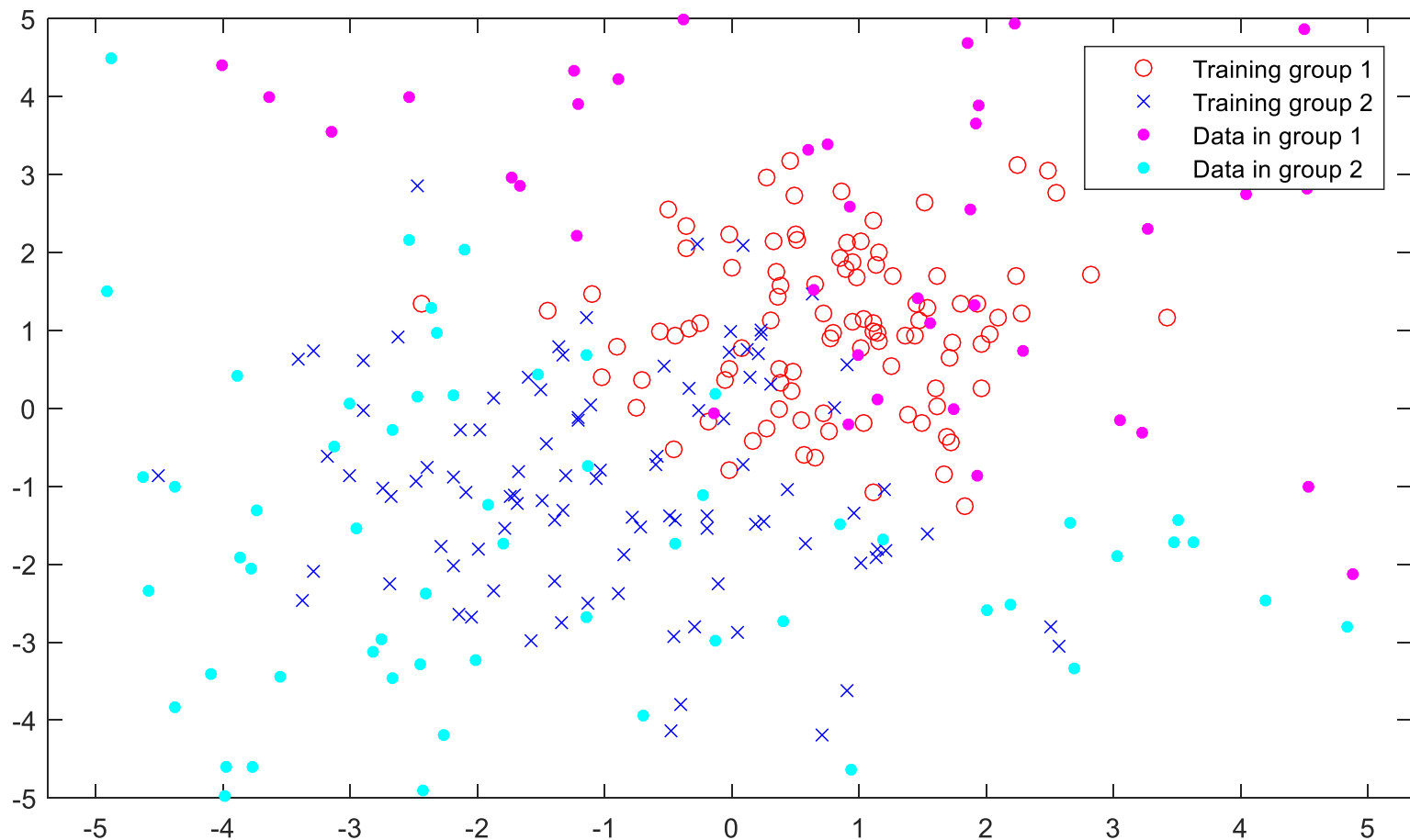


- Two sets of random data are generated.
- These are shown in the figure on the left hand side.
- The two groups are named as training group # 1 and group # 2

KNN classification

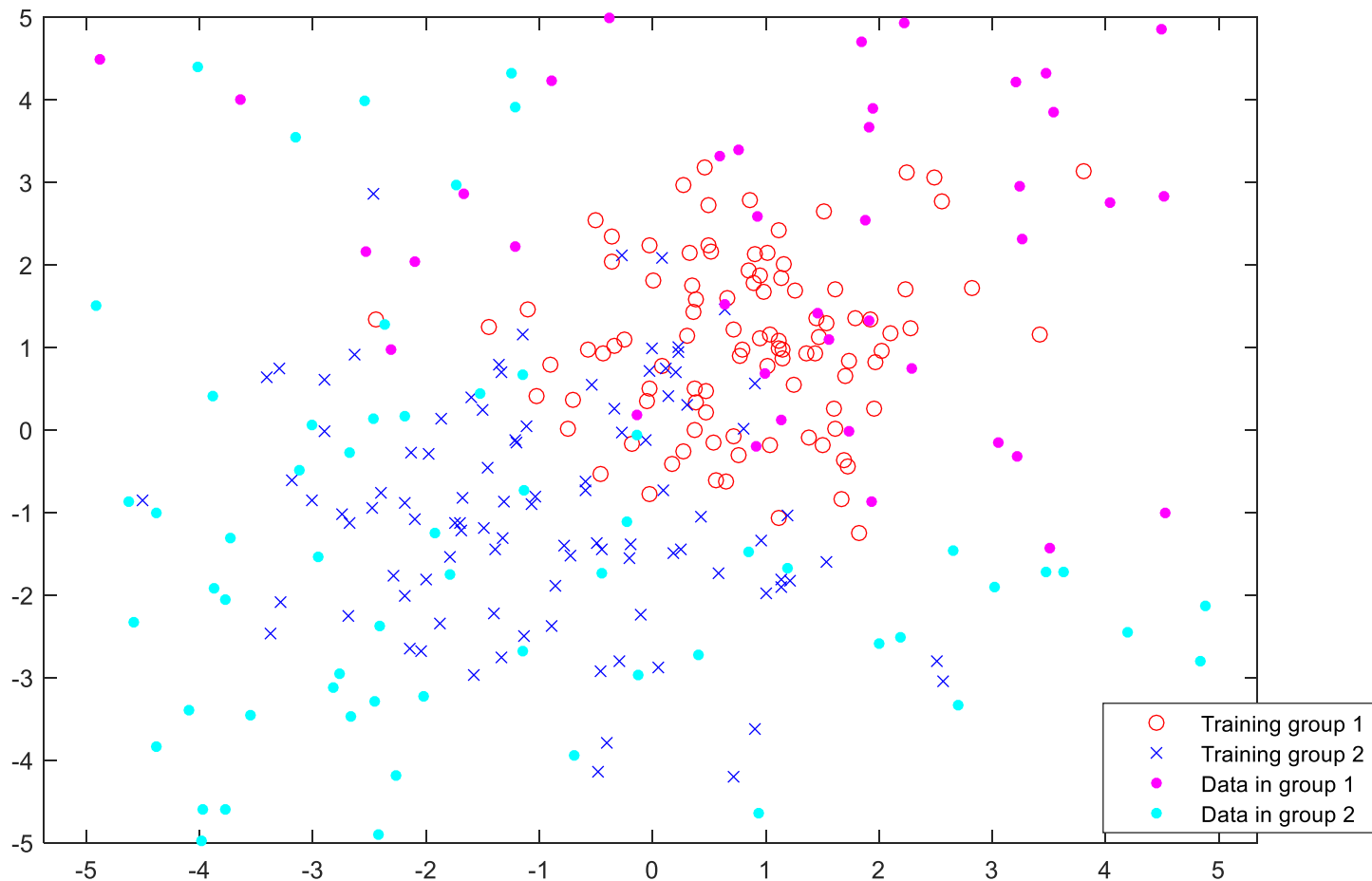


Classification



Classification

2 neighbors



K-Means Clustering

- K-means clustering can be used for understanding or mapping spatial and temporal clusters.

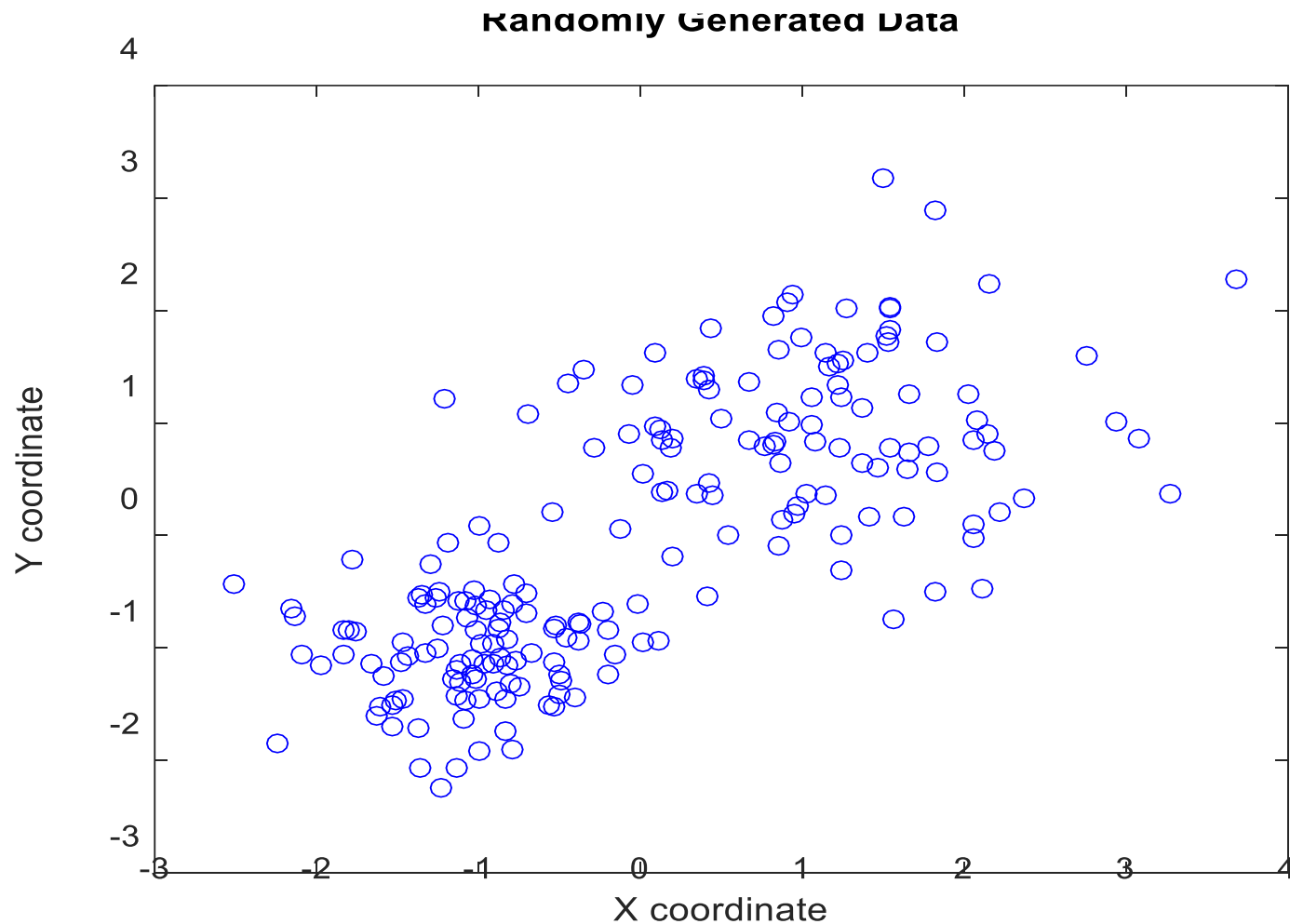
K-means Clustering Approach

- Clusters of data are obtained using a **K-means clustering** approach. K-means clustering method partitions the data into a number of clusters by using an iterative algorithm defined below:
- The generic K-means clustering algorithm is defined by the following steps:
 - Select K points as initial centroids
 - Repeat
 - Form k clusters by assigning each point to its closest centroid
 - Recomputed the centroid of each cluster
 - Until centroids do not change.

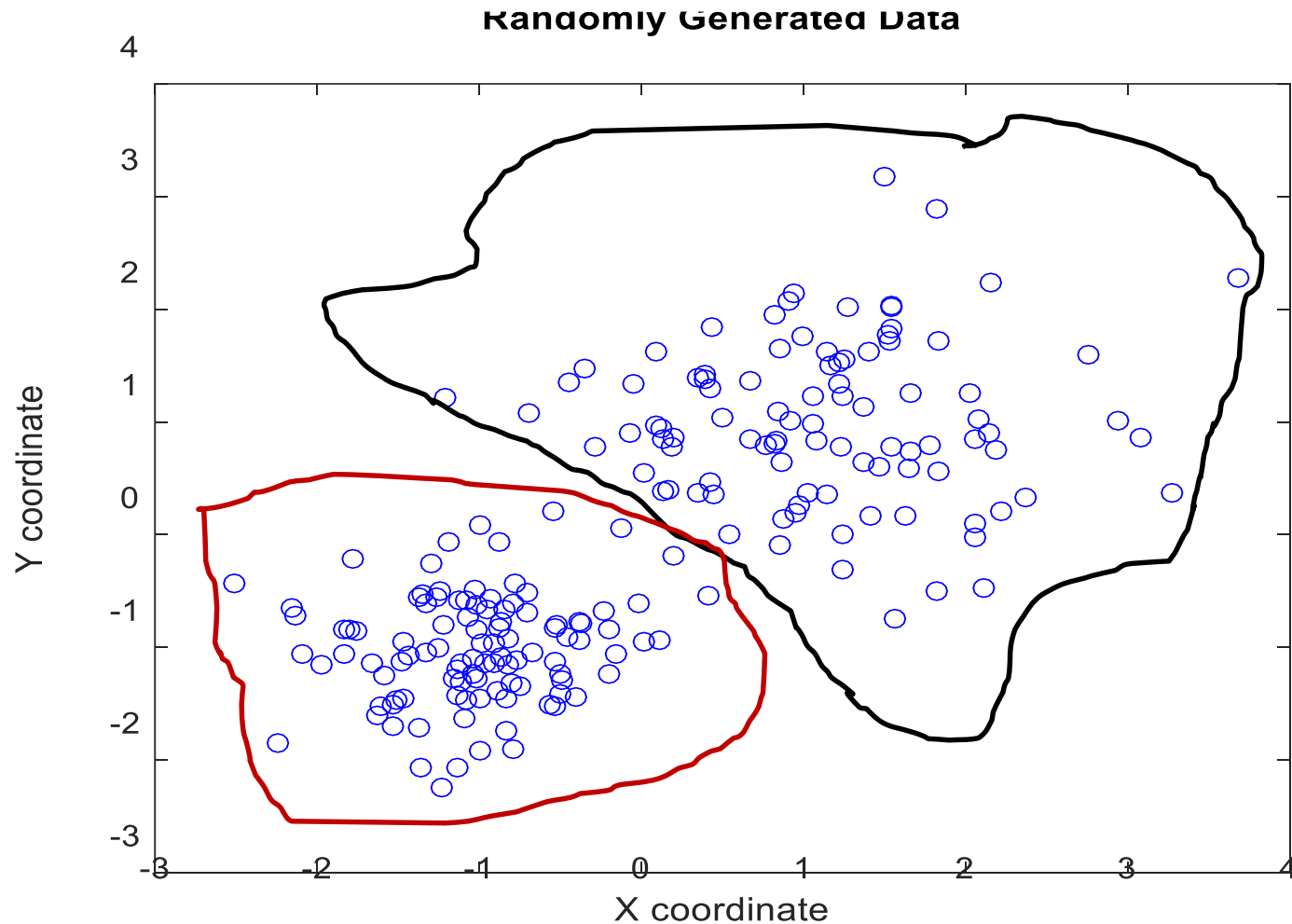
K-Means Clusters

- This algorithm is initially used to identify the clusters from the data and further processing is carried to identify the outliers.
- Initially k clusters are defined based on the data and then each observations (or data point) is evaluated for belongingness to one of the clusters based on its proximity defined by distance metrics (e.g. Squared Euclidean, City Block, Hamming or Correlation distance).
- The distance is calculated based using the center (centroid) of cluster and the observation. Once the process of assigning all the observations to all clusters is completed, the centers (i.e. centroids) are recalculated and all the observations are again assigned.
- The process of assignment and re-calculation of centroids is repeated iteratively till there is no change in the centroid values.
- When no change is noted in the centroid values or the differences of values from the previous iteration are smaller than a pre-defined threshold value, convergence is achieved and the execution of algorithm is terminated.

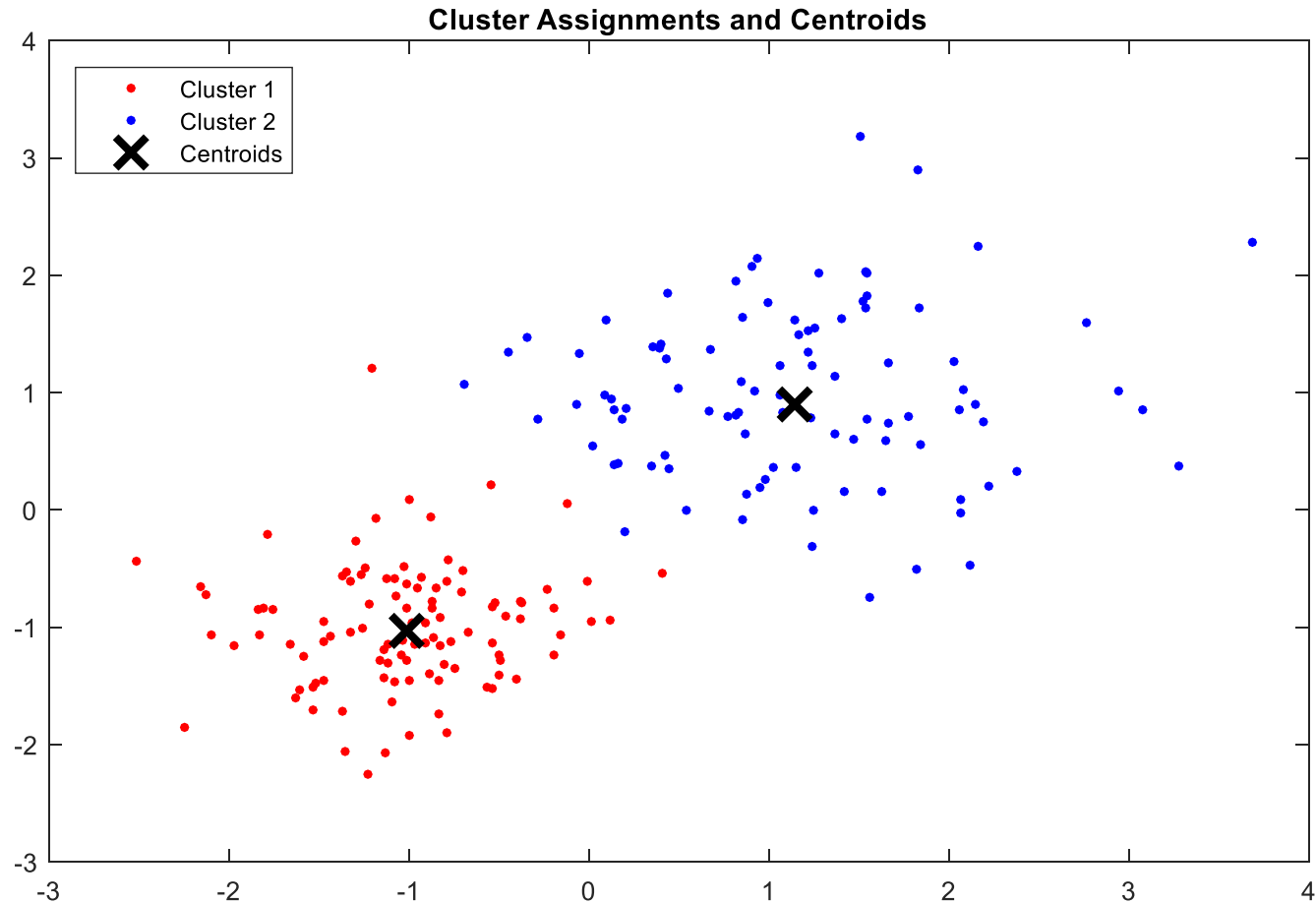
Random data in space



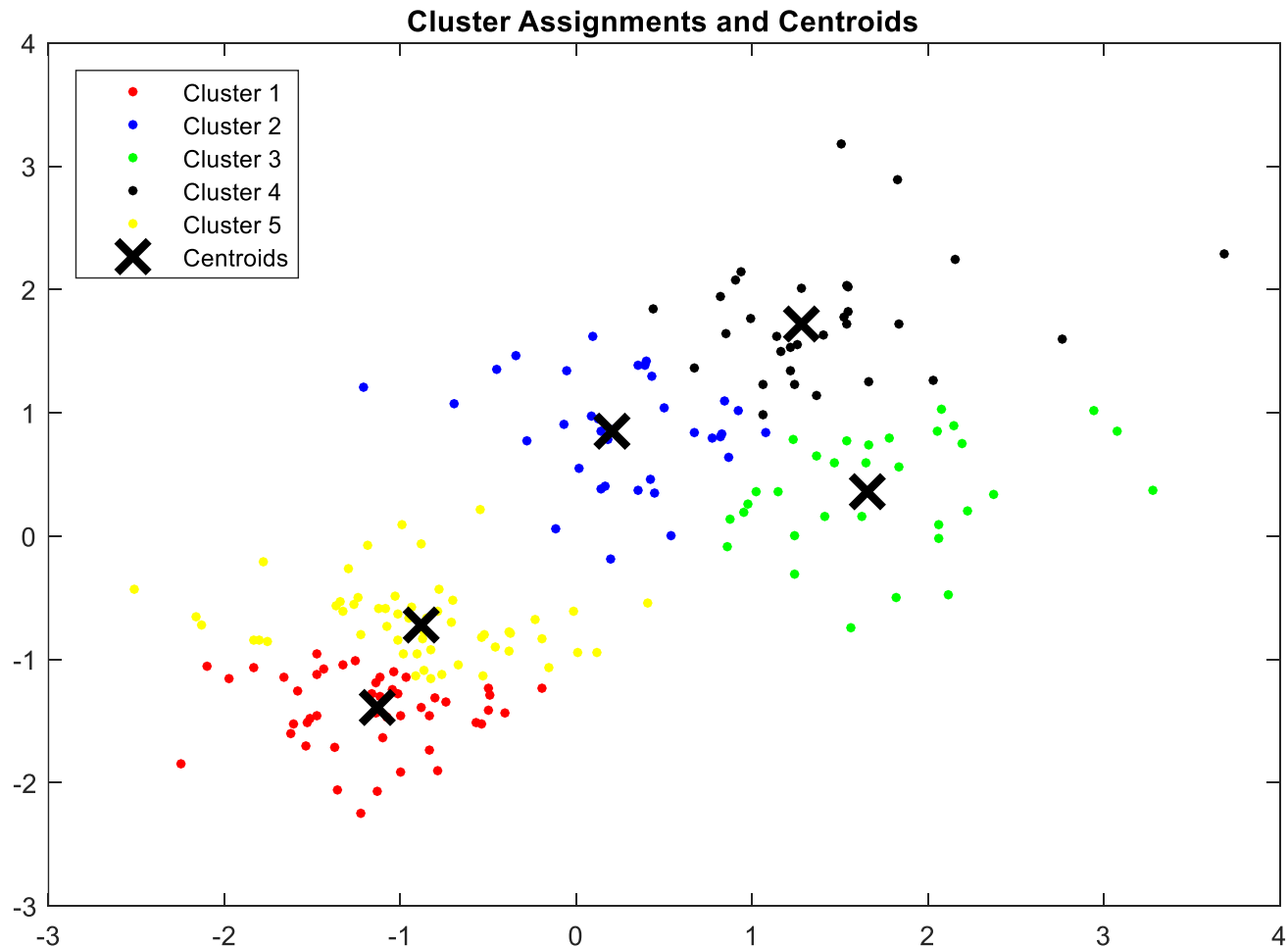
Visual identification of clusters



Two clusters using K-means



Five clusters identified using k-means



Example of K-means Clustering Application

- Identification of outliers in stage data (i.e. Water level data)
- Application of K-means clustering for temporal data.

Data Cleaning

- **Data cleaning** is one of the first steps in data storage and analysis process requiring identification of outliers, non-homogeneous observations and datasets suspected to be influenced by instrumental and sensor-based, human and transcription errors.
- Hydrologic and climate data measured under varying field conditions and multiple sensors are known to be plagued by the problem of data **outliers and anomalies**.
- Techniques for identifying outliers and methods for performance evaluation of anomaly detection methods are critical for task of maintaining **unbiased, clean and error-free homogeneous data**.

Terminology

- Outliers,
- Anomalies,
- Extreme observations,
- Contaminants,
- Exceptions,
- Excursions,
- Noise.

Visual Evaluation

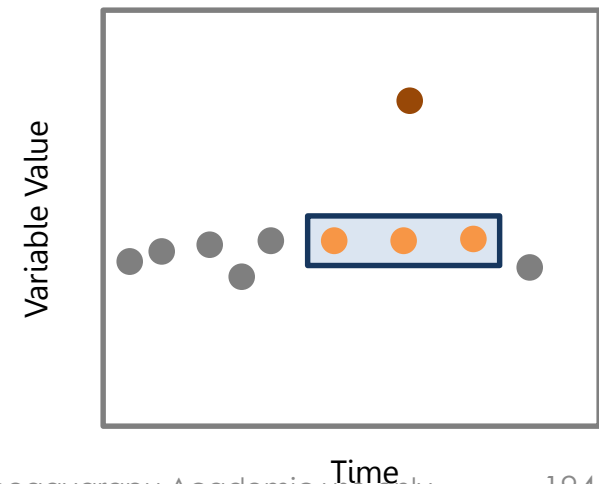
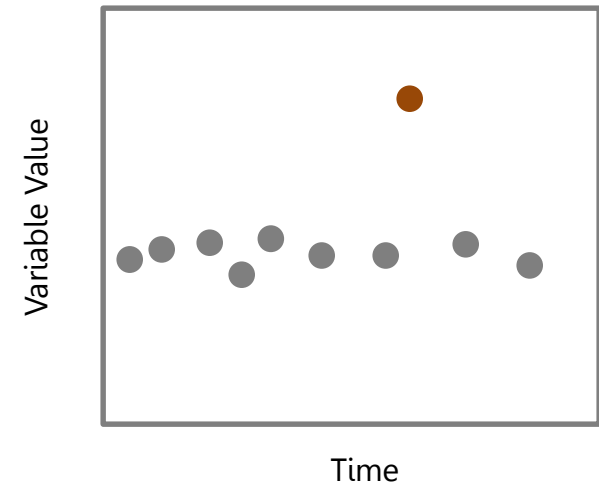
- Visual detection of outliers is possible in many instances
- More subjective, depends on the perception of the analyst in order to determine the values that are "very far away" and "low frequency."
- Manual inspection of data is also an extremely time-consuming task and, not suitable for most automatic large hydro-monitoring networks.

Outlier and Anomalies

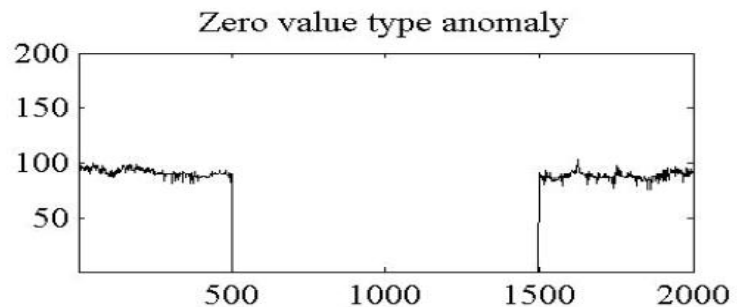
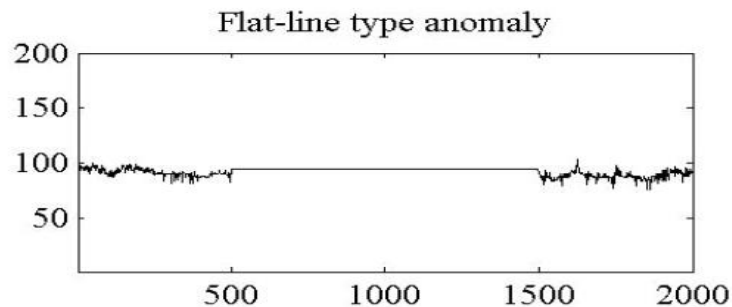
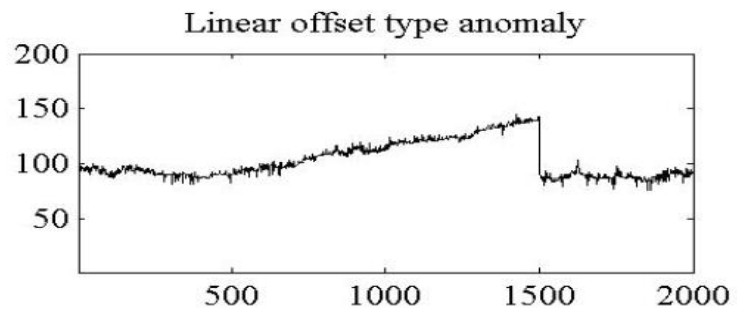
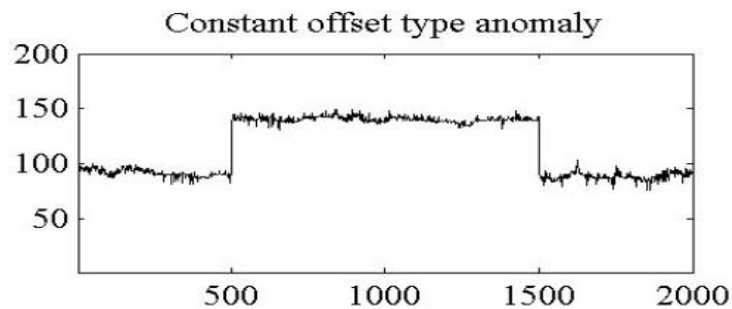
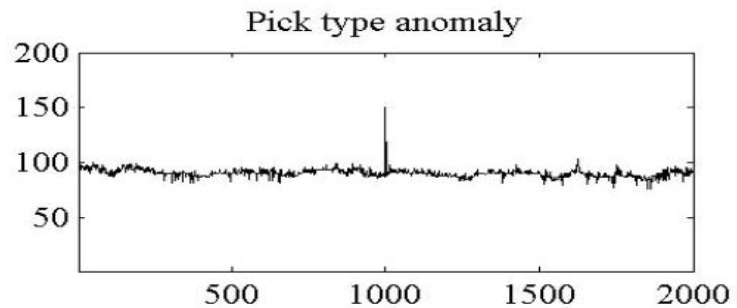
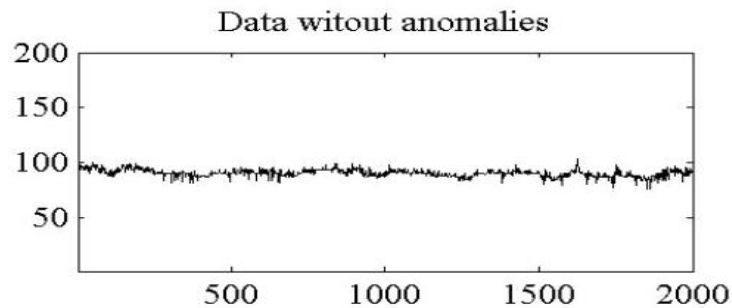
- The words “outlier” and “anomaly” are used interchangeably in many studies.
- An outlier is an unusual observation, numerically different from a set of observations.
- Anomaly refers to a pattern in a given data set that does not conform to an established normal behavior.
- Anomalies are more difficult to detect with statistical methods. Rule-based methods derived from expertise will help immensely.
- Outliers are also referred to as **Distributional Monsters**.

● Outlier

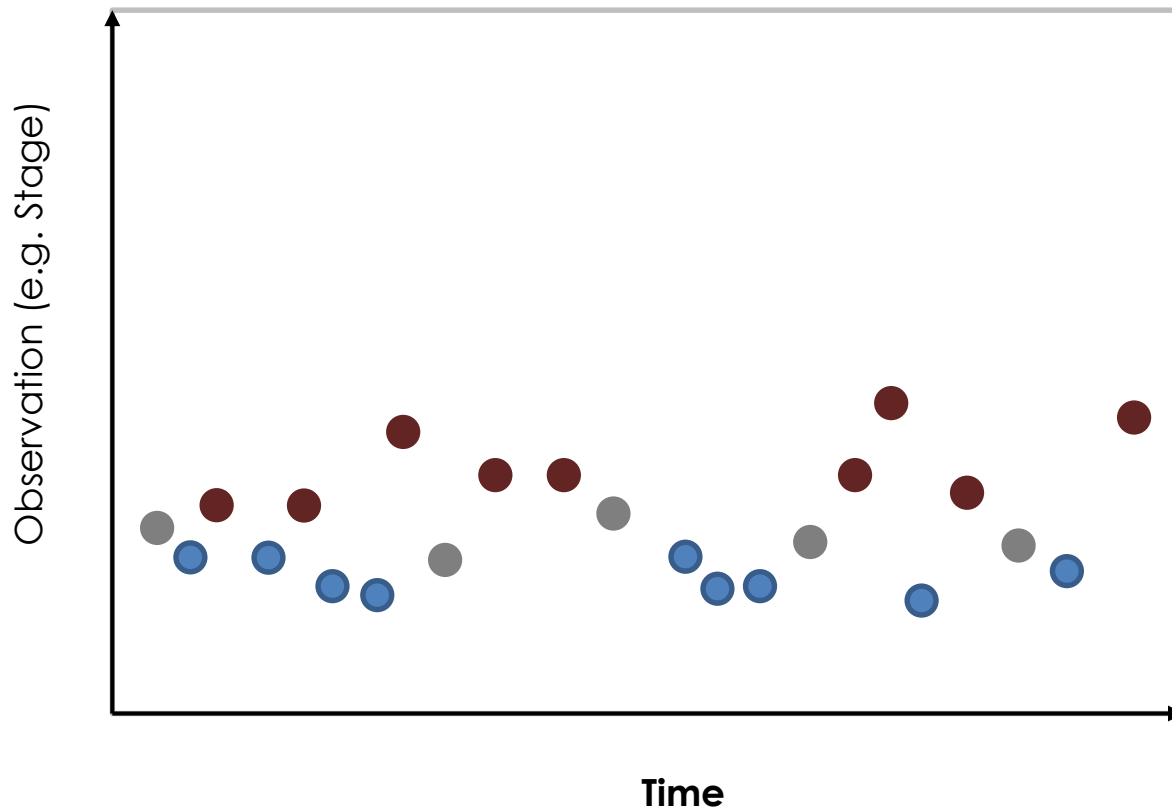
● ● ● Anomalous observations



Data Anomalies



Clustering Approach

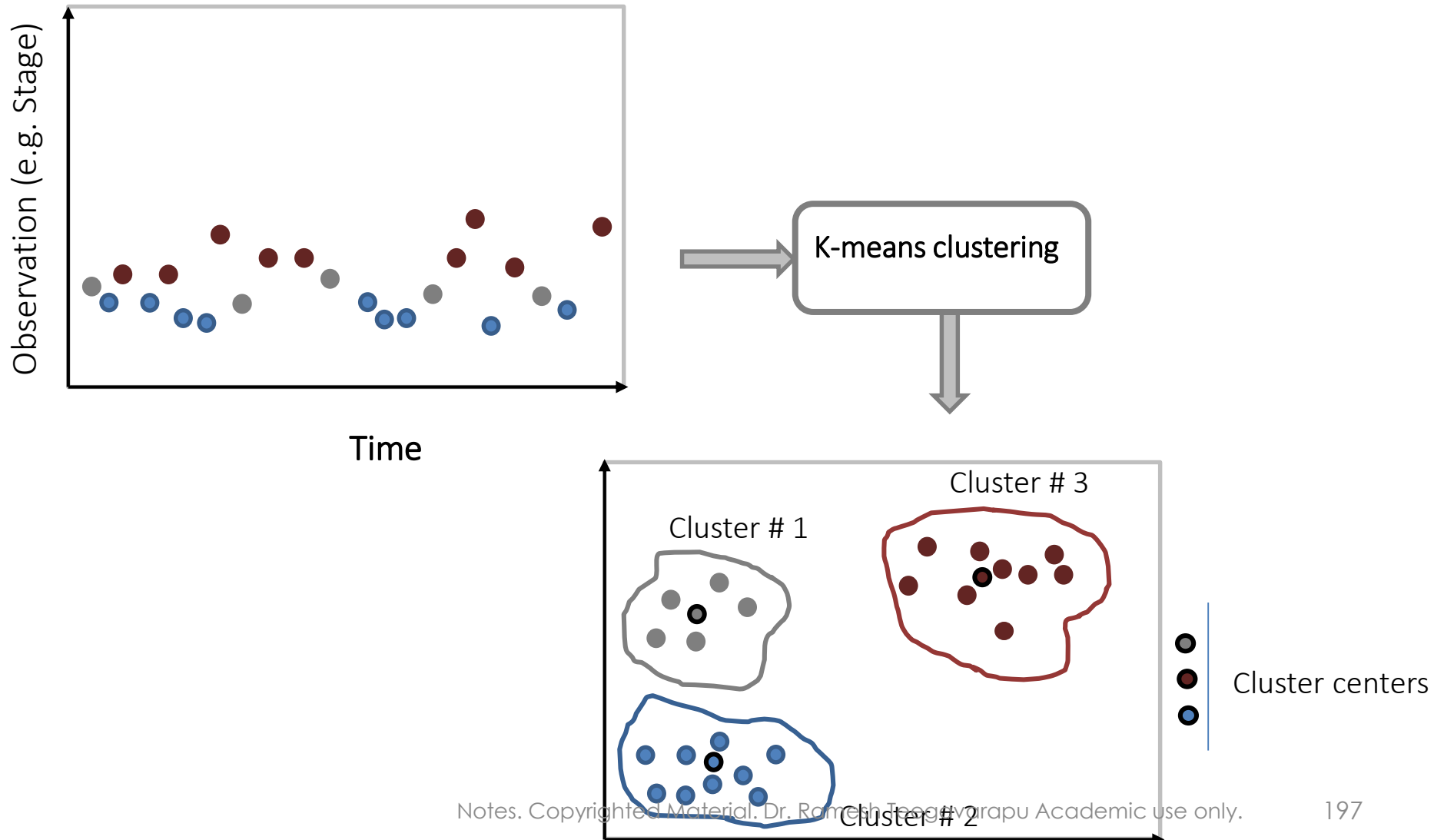


Outlier
Identification
using clusters

Identification of Clusters

K-means

Clustering



Clustering approach for stage data

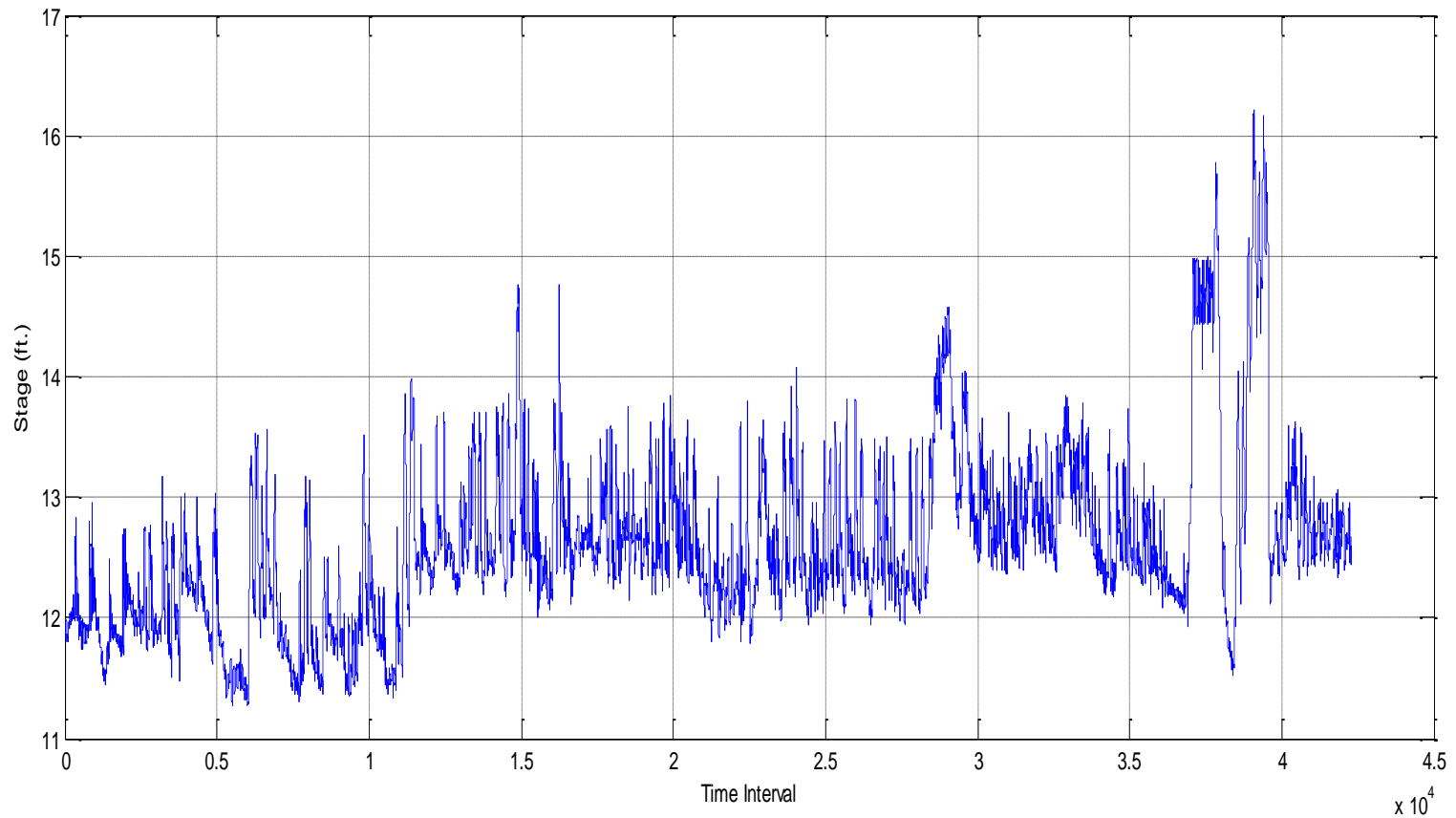
- The following steps are involved in identification of outliers in stage data using the K-means-based approach:
 - Obtain the time series of stage data
 - Specify the number of clusters (k) (this is usually user specified or can be a number greater than a specific threshold value)
 - Evaluate each cluster and select the cluster (K^s) that contains the centroid with the highest numerical value
 - Obtain the 95th (Q95) or 99th (Q99) percentile value based on all the observations in the cluster K^s .
 - Select all the observations in the cluster K^s and flag all the observations that are larger than Q95 or Q99 threshold value.

-
- A variant of the above method can be used in which the outliers are identified in the selected cluster using Tukey's Box-Plot or Adjusted Box Plot method instead of using Q95 or Q99 threshold values.
 - Alternatively, the centroid or the median value of each cluster can be used as a threshold value to identify the outliers.

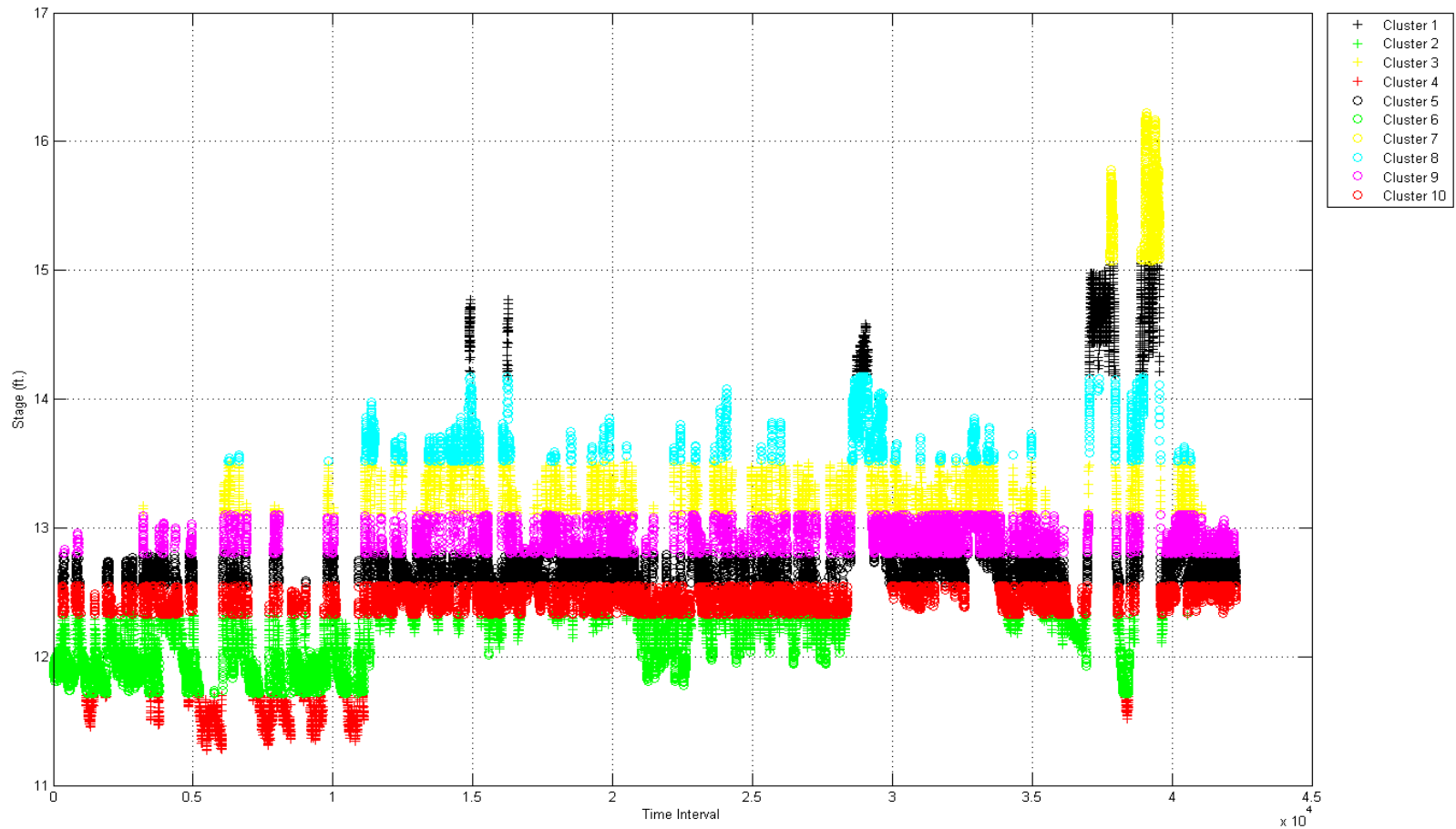
Application Example (stage outlier application)

- A total of 10 clusters are specified and based on the dataset of 42,278 observations, the centers of the clusters identified by numerical values of stage (in feet) with the help of K-means clustering algorithm are: 14.611, 12.205, 13.288, 11.53, 12.669, 11.884, **15.524**, 13.748, 12.93, 12.443.
- The cluster with the highest centroid value is selected and in the current context it is the centroid with numerical value of 15.524 ft. (seventh cluster).
- All the 10 clusters identified by the k-means approach. By selecting the clusters with lowest and highest centroids, lower end and higher end outliers can be identified.

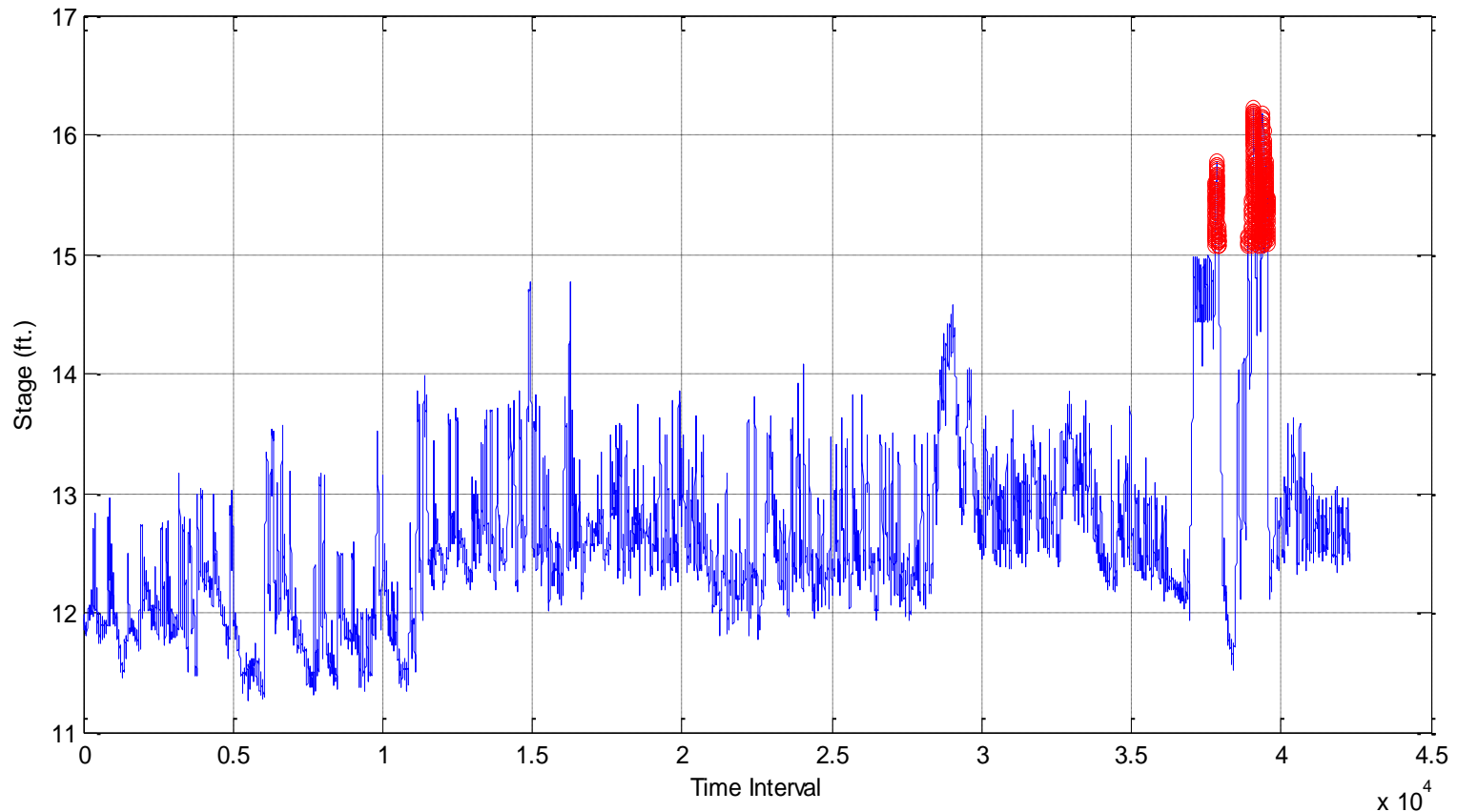
Example (stage data)

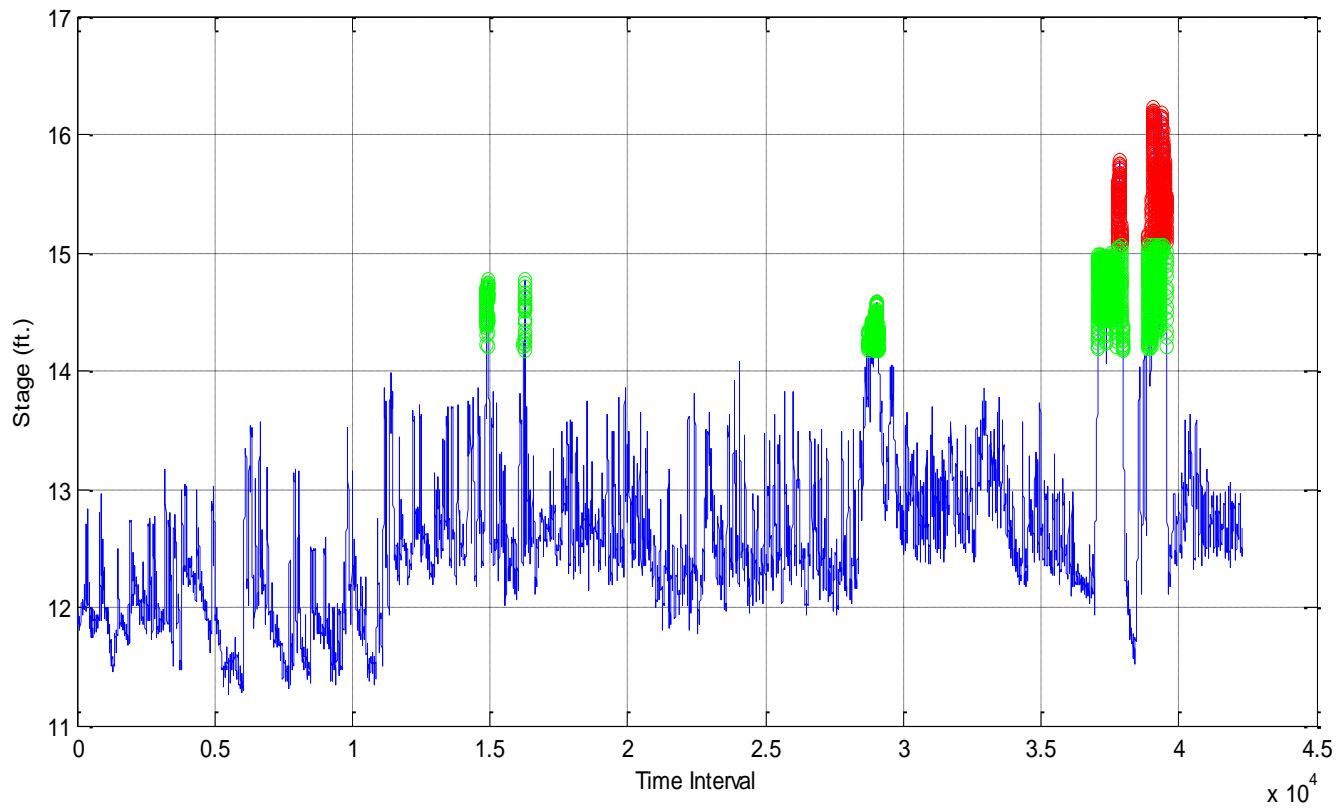


10 Clusters.

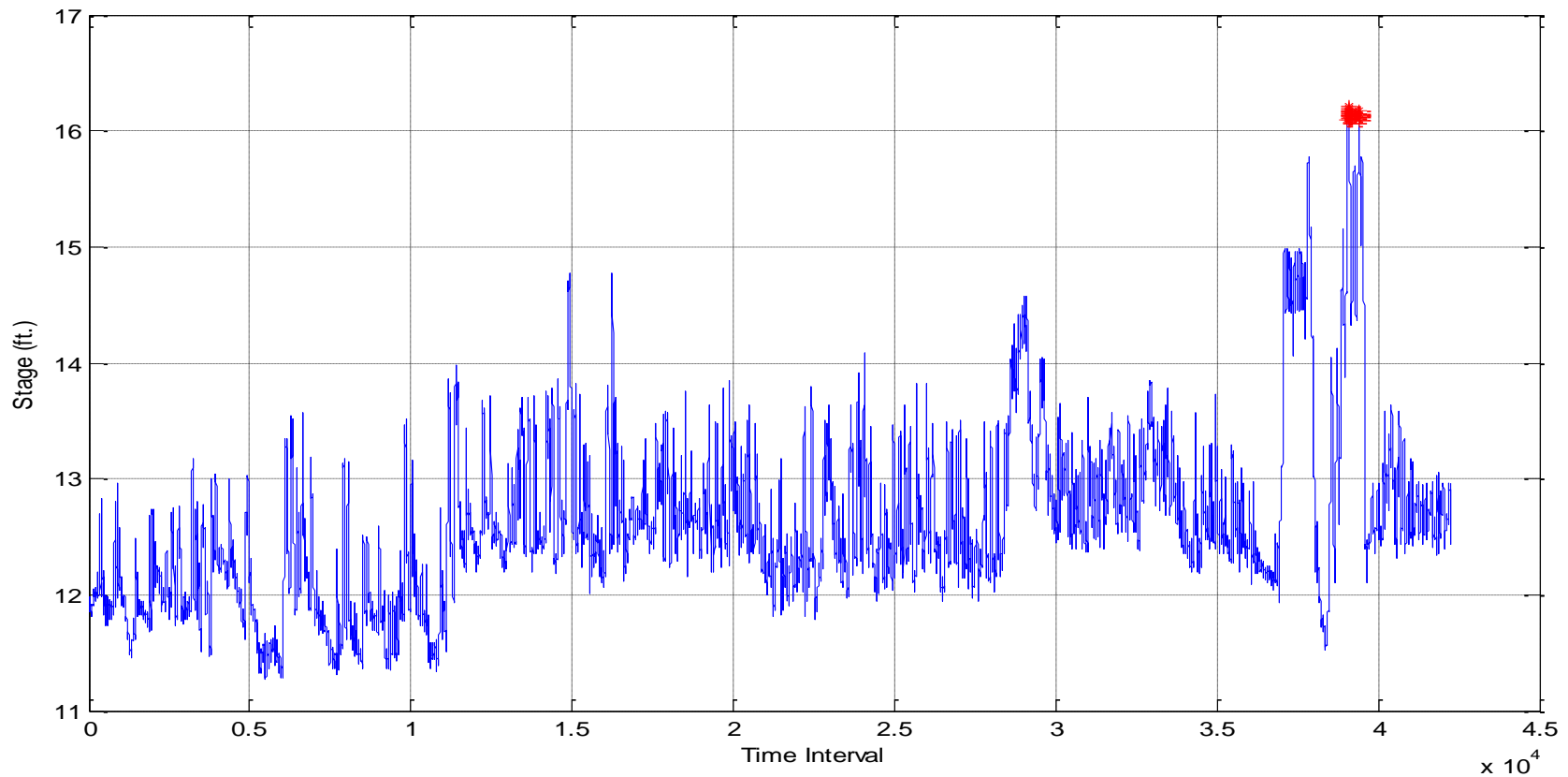


Data from the cluster (max median value)

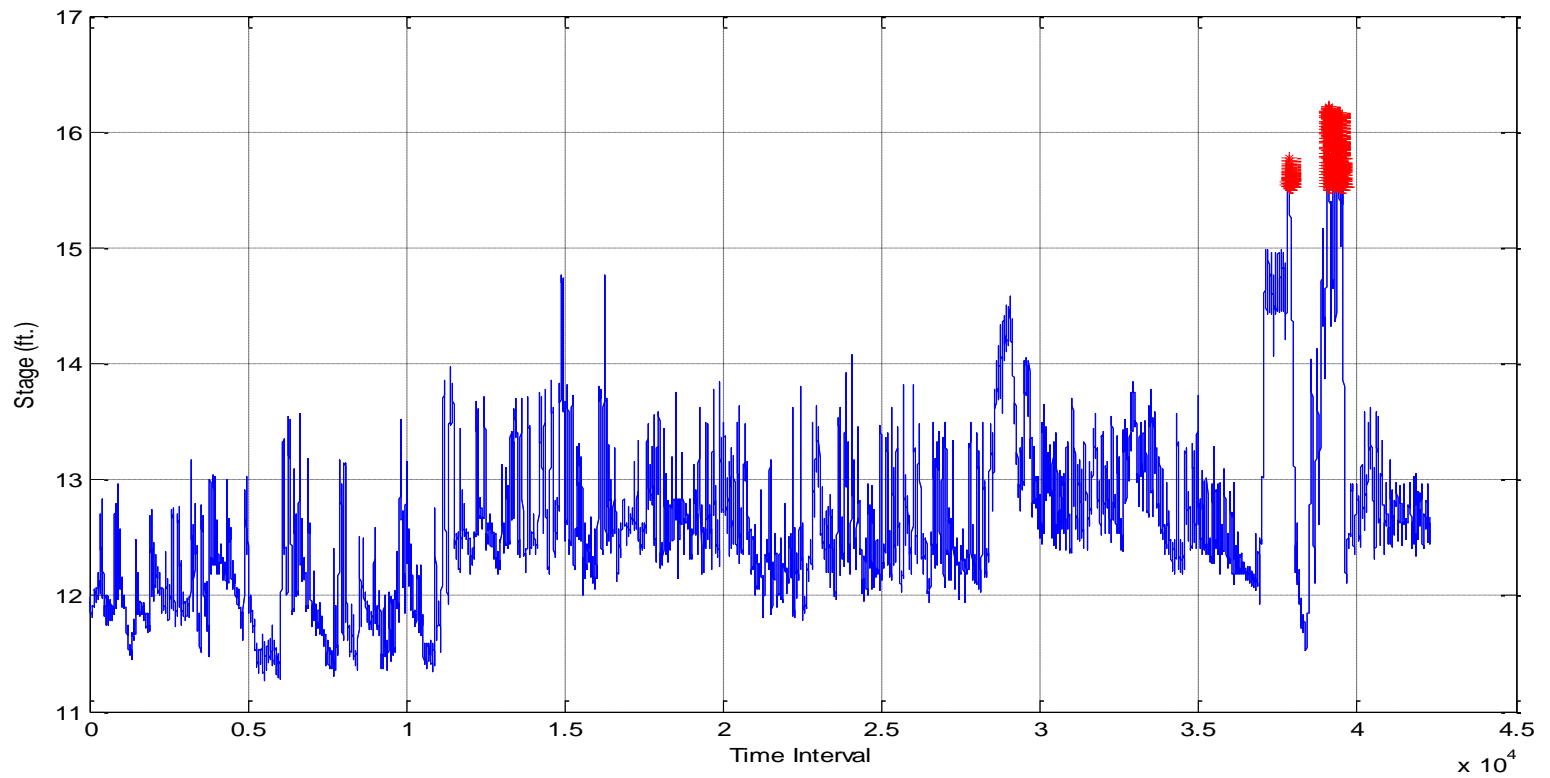




Observations >Q95



Observations



Issues

- A number of issues that arose when the evaluation of the cluster-based method that need to be resolved are:
 - Decision regarding the **number of clusters**.
 - Criteria for selection one or more clusters for further processing.
 - Selection of **thresholds for identification of outliers**.
 - Incorporation of changes to the thresholds and selection process of clusters based on domain knowledge and the any information gained from historical data.
 - Additional outliers processing methods to be included in conjunction with K-means clustering method.

Spatial Autocorrelation

- In many instances we need information about the location of point instances and their attributes to decipher spatial patterns.
- The spatial autocorrelation measure provides the idea of clustered/dispersed the point locations are with respect to the attribute values.

Tobler's first law of geography (1970) :

- *“Everything is related to everything else, but near things are more related than distant things”,*

which forms the basis for many interpolation techniques.

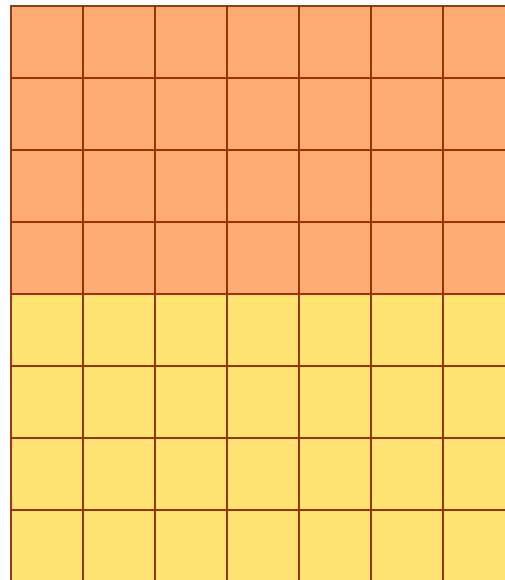
Spatial autocorrelation

- Spatial autocorrelation of a set of points refers to the degree of similarity between points or events occurring at these points and points or events in nearby locations.
- If positive correlations exist in a point distribution, points with similar characteristics tend to be near each other.
- Alternatively if spatial autocorrelation is weak or nonexistent, nearby patterns do **not exhibit** any similar or dissimilar pattern or a random pattern exists. – Refer to “First Law of Geography”

Spatial Autocorrelation Coefficient (SAC)

- SAC – we can measure
 - Proximity of locations
 - The similarity of the characteristics of these locations
- Two measures are routinely used to obtain SAC
 - Geary's Ratio C
 - Moran's I Index.

Positive Spatial Autocorrelation



Joint Counts

SAC

S: represents the similarity, W
represents proximity

$$SAC = \frac{\sum_{i=1}^n \sum_{j=1}^n S_{ij} W_{ij}}{\sum_{i=1}^n \sum_{j=1}^n W_{ij}}$$

Geary's Ratio C (contiguity)

$$C = \frac{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - x_j)^2}{2 \sum_{i=1}^n \sum_{j=1}^n W_{ij} S^2}$$

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

Moran's I index

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sigma^2 \sum_{i=1}^n \sum_{j=1}^n W_{ij}}$$

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

Numeric Scales and explanations

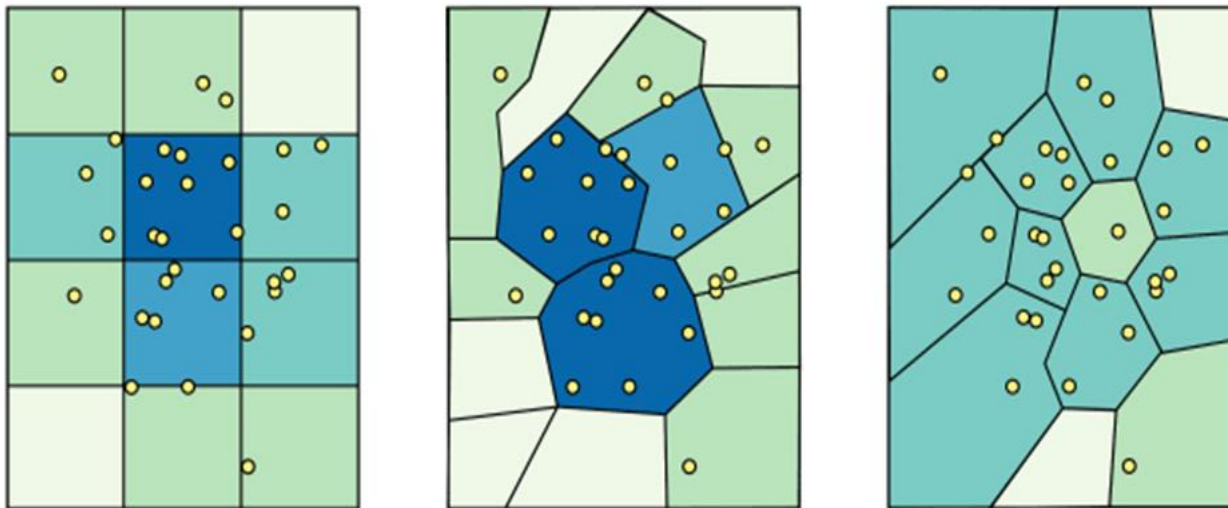
- Clustered pattern in which adjacent or nearby points show similar characteristics
 $0 < C < 1$
- Random pattern in which points do not show particular patterns of similarity
 $C \sim 1$
- Dispersed pattern in which adjacent or nearby points show different characteristics
 $1 < C < 2$

Modifiable Areal Unit Problem (MAUP)

- MAUP is a major **contentious issue** in spatial analysis.
- The MAUP is “a problem arising from the **imposition of artificial units of spatial reporting on continuous geographical phenomena resulting in the generation of artificial spatial patterns**” (Heywood, 1988). In other words, artifacts or errors are created when one groups data into units for analysis.

MAUP

Scale (or aggregation) and zone (or grouping).



Source of figure: <http://gispopsci.org/maup/>

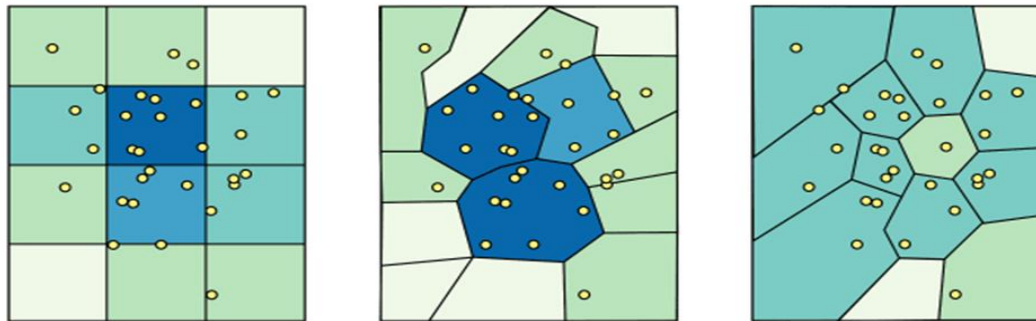
Notes. Copyrighted Material. Dr. Ramesh Teegavarapu Academic use only.

Scale

- The scale at analysis to be conducted should be carefully chosen.
- For each scale chosen, the results produced are completely different.
- For any analysis it is important to choose your scale to match your research question.
- In some instances, coarser or finer scale may be better to answer a question.
- Finer-scale data can be aggregated, while coarser scale data cannot easily be divided.

Zone **MAUP**

- Dividing a spatial unit or region into several zones without clear understanding of the process or the main question investigated can be a problem.



- Examples: Gerrymandering or Political redistricting.

The Pitfalls and Potential of Spatial Data

Independent variable Dependent variable

87	95	72	37	44	24
40	55	55	38	88	34
41	30	26	35	38	24
14	56	37	34	8	18
49	44	51	67	17	37
55	25	33	32	59	54

72	75	85	29	58	30
50	60	49	46	84	23
21	46	22	42	45	14
19	36	48	23	8	29
38	47	52	52	22	48
58	40	46	38	35	55

Aggregation scheme 1

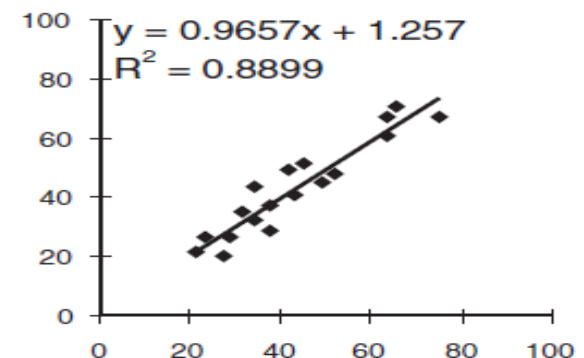
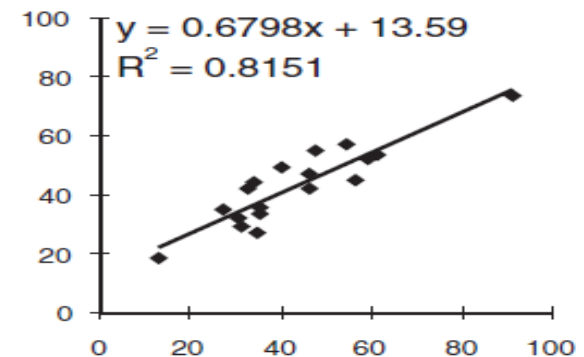
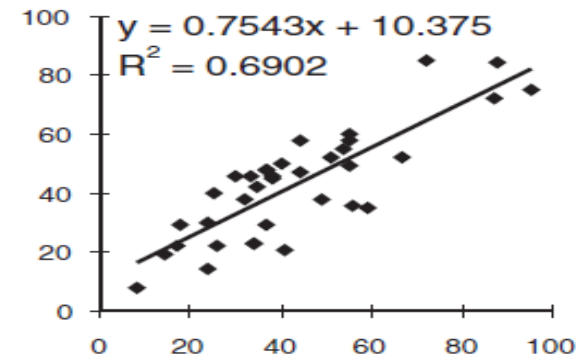
91	54.5	34
47.5	46.5	61
35.5	30.5	31
35	35.5	13
46.5	59	27
40	32.5	56.5

73.5	57	44
55	47.5	53.5
33.5	32	29.5
27.5	35.5	18.5
42.5	52	35
49	42	45

Aggregation scheme 2

63.5	75	63.5	37.5	66	29
27.5	43	31.5	34.5	23	21
52	34.5	42	49.5	38	45.5

61	67.5	67	37.5	71	26.5
20	41	35	32.5	26.5	21.5
48	43.5	49	45	28.5	51.5



Source: Example of MAUP from Unwin, 2009, Wiley.

Ecological Fallacy

- **Ecological fallacy**—The ecological fallacy occurs when **one mistakenly assigns a statistic or value that has been calculated for a group to a member of that group.**
- For example, if one knows that residents of a certain area have an average income well over \$50,000 per year, it does not mean that an individual from that area necessarily makes more than \$50,000 each year.

Smoothing of Spatial Data

- In many instances, spatial data needs to be smoothened due to presences of noise or other issues.
- Spatial Filters are used to smoothen the data.
- Examples of spatial filters include:
 - Mean
 - Median
 - Other variants (Gaussians etc.)
- Mean and median filters are used commonly in geosciences and also in image processing.

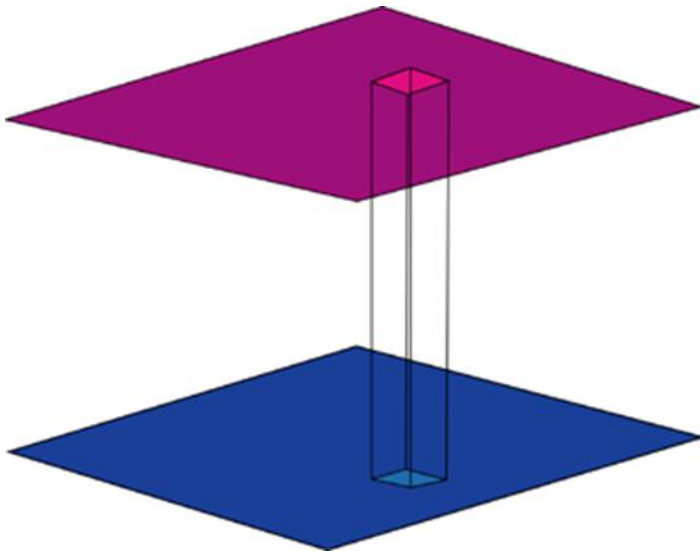
Spatial Filters

- Spatial filter is a form of **focal operator**, whereby the results for a given cell are a function of the value of the neighboring cells

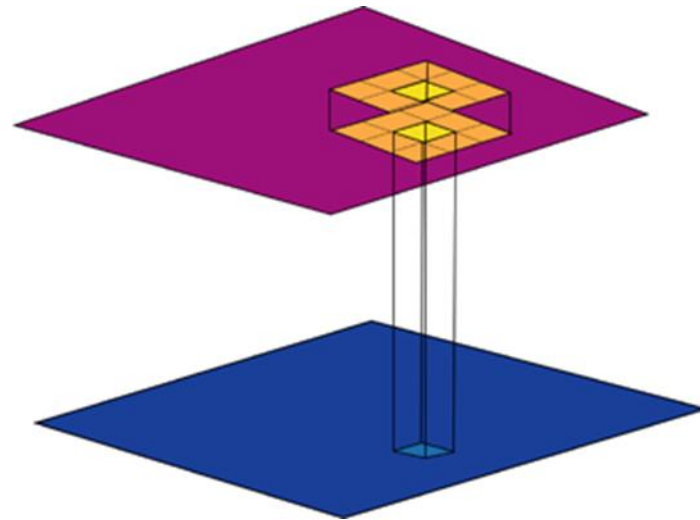
Functions

- Single cell locations (**local** functions)
 - Only one cell is used*
- Cell locations within a neighborhood (**focal** functions)
 - Mean and Median Filters
- Cell locations within zones (**zonal** functions)
 - Cells are used that lie in a zone.
- All cells within the region (**global** functions)

Local and Focal Operators

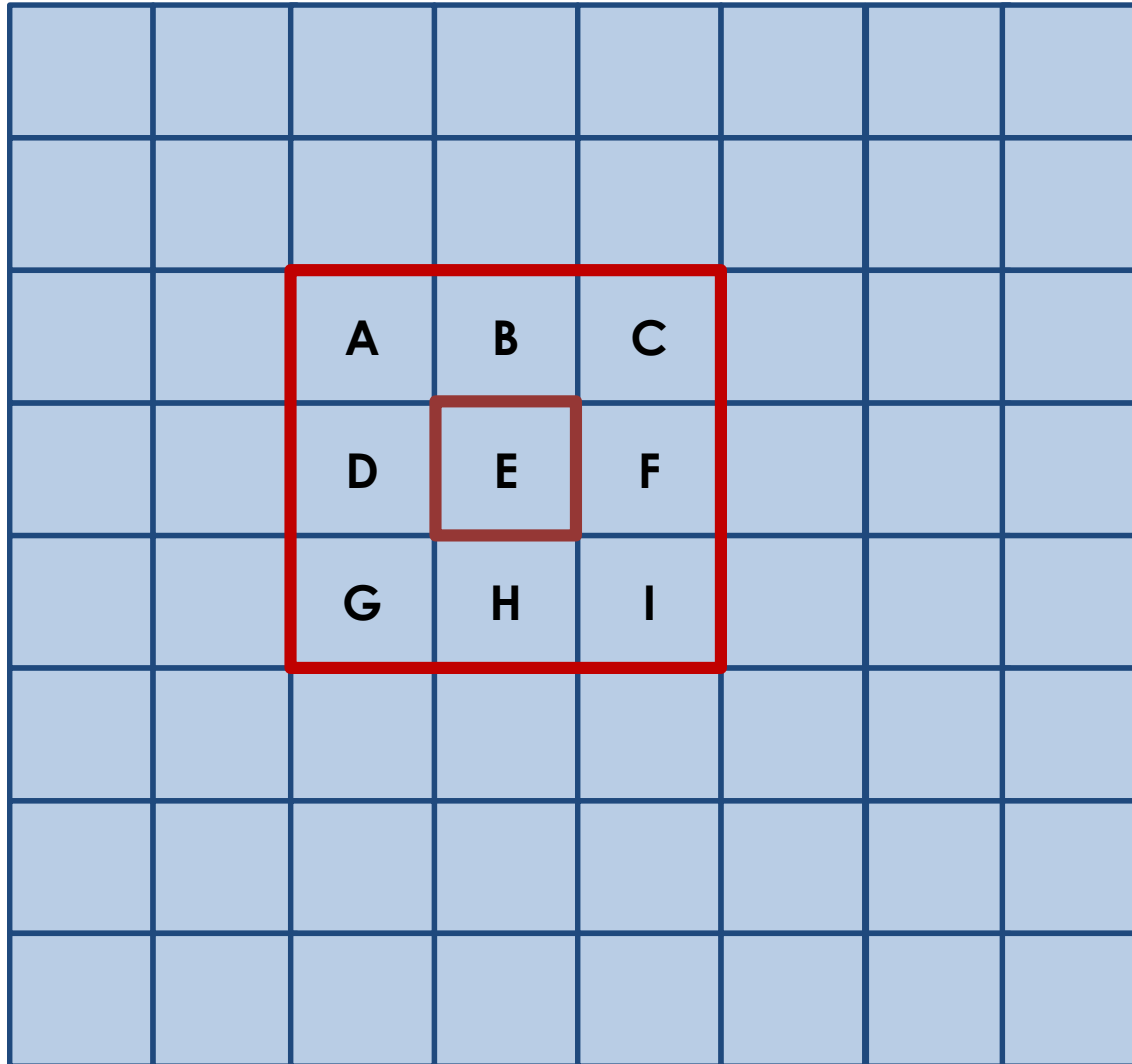


Local



Focal

Mean Filter



The value of cell
Is replaced by
Mean of all the
Values
surrounding
It. Including the
cell
Value under
consideration
3 X 3

Mean Filter

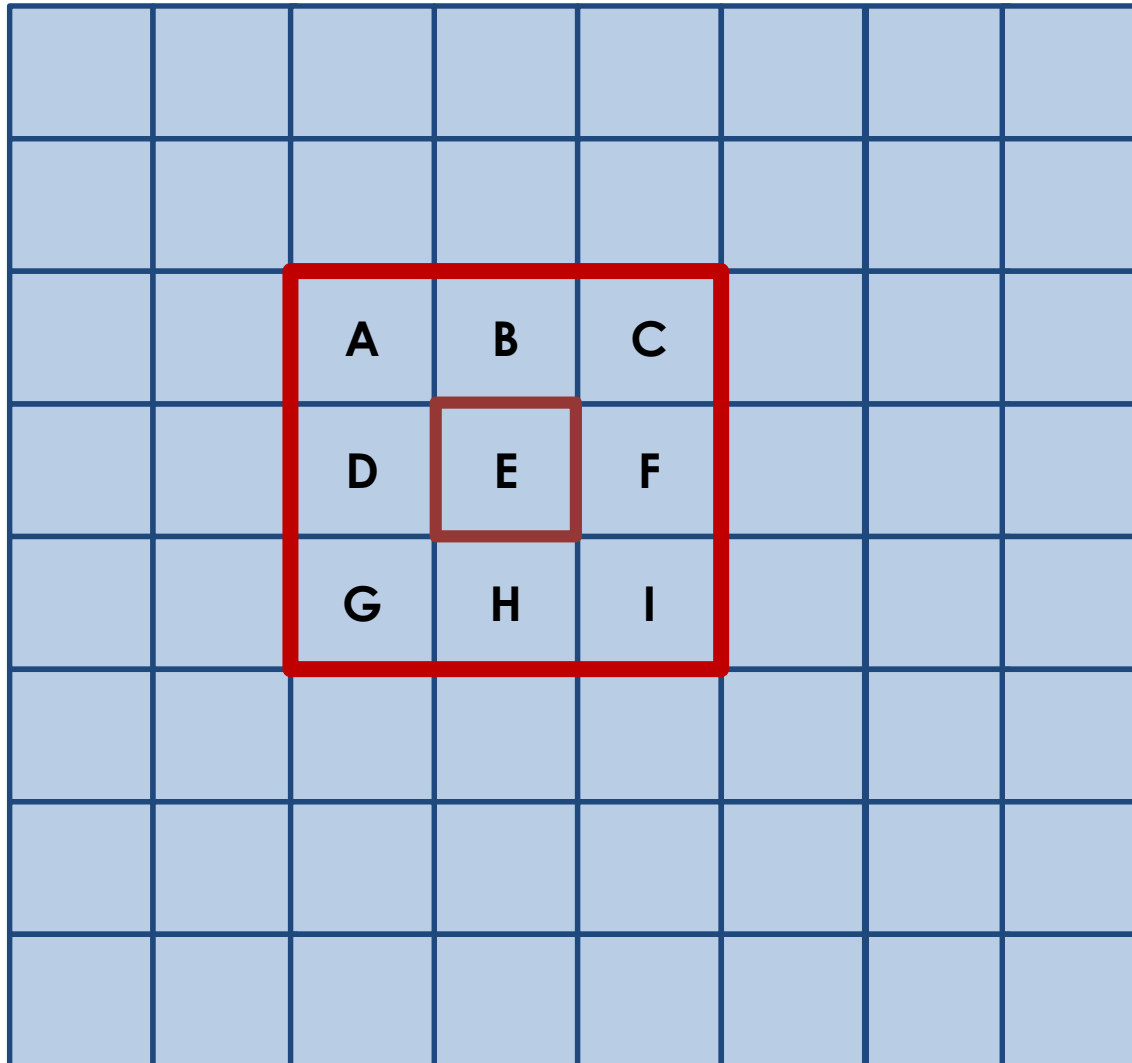
4	5	0	3
2	3	4	7
6	1	8	4
3	9	2	0

A.

3.50	3.00	3.67	3.50
3.50	3.67	3.89	4.33
4.00	4.22	4.22	4.17
4.75	4.83	4.00	3.50

B.

Median Filter



The value of cell
Is replaced by
median of all the
Values
surrounding
It. Including the
cell
Value under
consideration
3 X 3

The region size
can be increased to
5 X 5 cells also.

Filters

- The median and mode are also sometimes computed in a moving window.
- The key advantage of the median is that it is not sensitive to outliers, unlike the mean. The mode can be derived from categorical data.
- The average is computed only from pixels in the neighborhood that have properties similar to the pixel at the center of the window.

Example: Trevi Fountain



Application of Mean Filter for an image

- Artificially add noise.
 - Salt and Pepper noise
 - Black and white pixels are flipped.
- Remove the noise using mean filter using 3 x 3 average kernel used.

Image with noise and filtered image

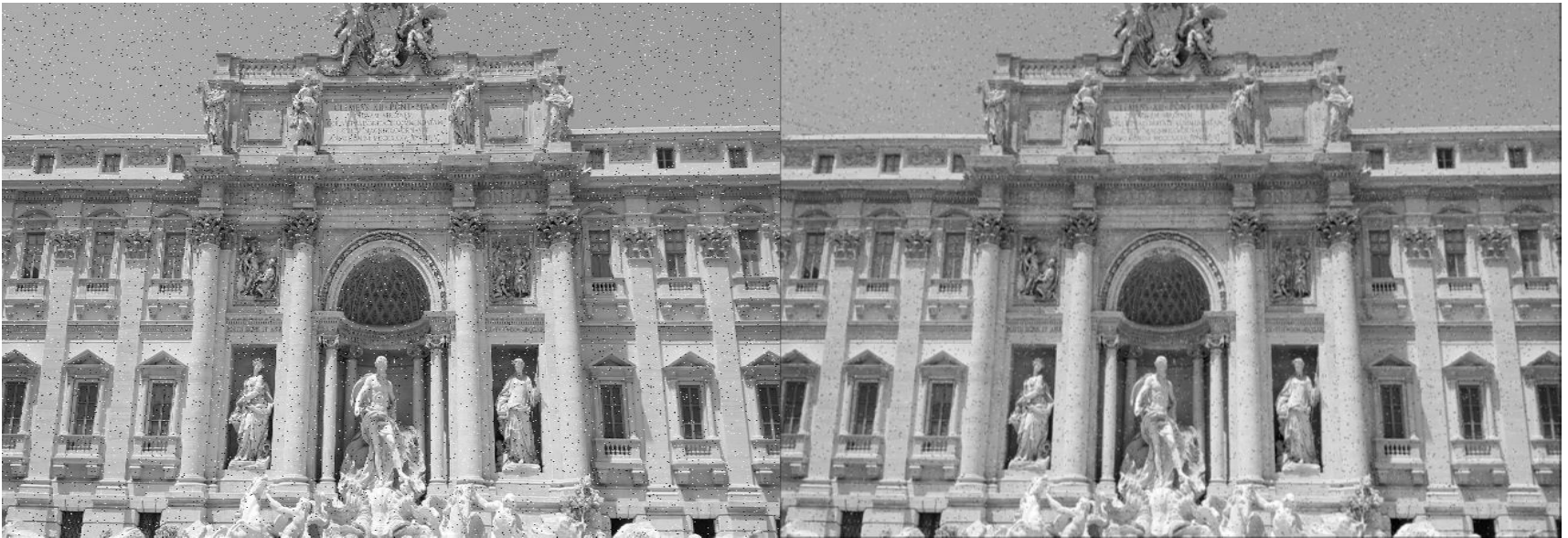


Image with noise

Image filtered

Filtered Image



Application of Median Filter for an image

- Artificially add noise.
 - **Salt and Pepper** noise
 - Black and white pixels are flipped.
 - Salt-and-pepper noise, for which a certain amount of the pixels in the image are either black or white (hence the name of the noise)
- Remove the noise using median filter using 3 x 3 average kernel.
- The median filter allows a great deal of **high spatial frequency detail** to pass while remaining very effective at removing noise

Image with noise added



Image with noise and filtered noise image



Image with noise

Image filtered

With Noise added

High level
of noise



Median filtered



Mean Filtering



Image with noise



Image filtered

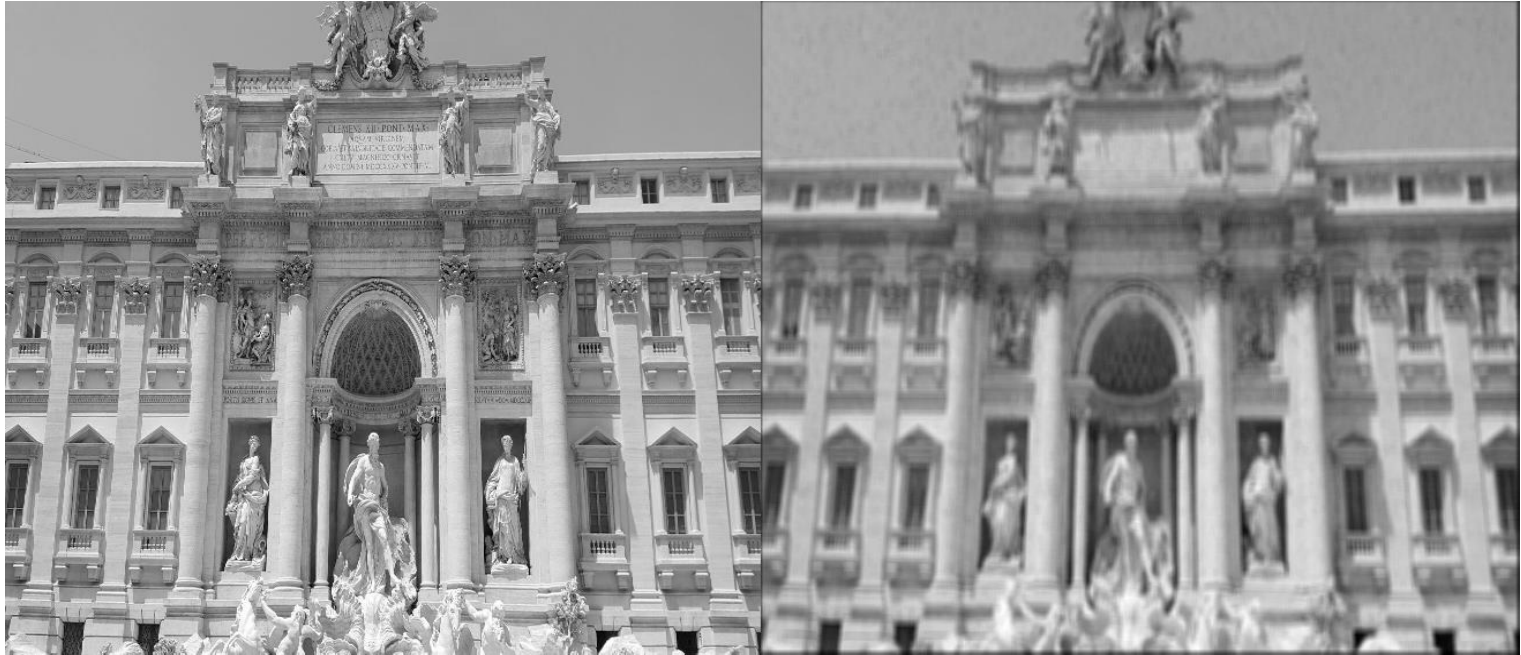
Introduces blurring

Filtered image

Mean filter used



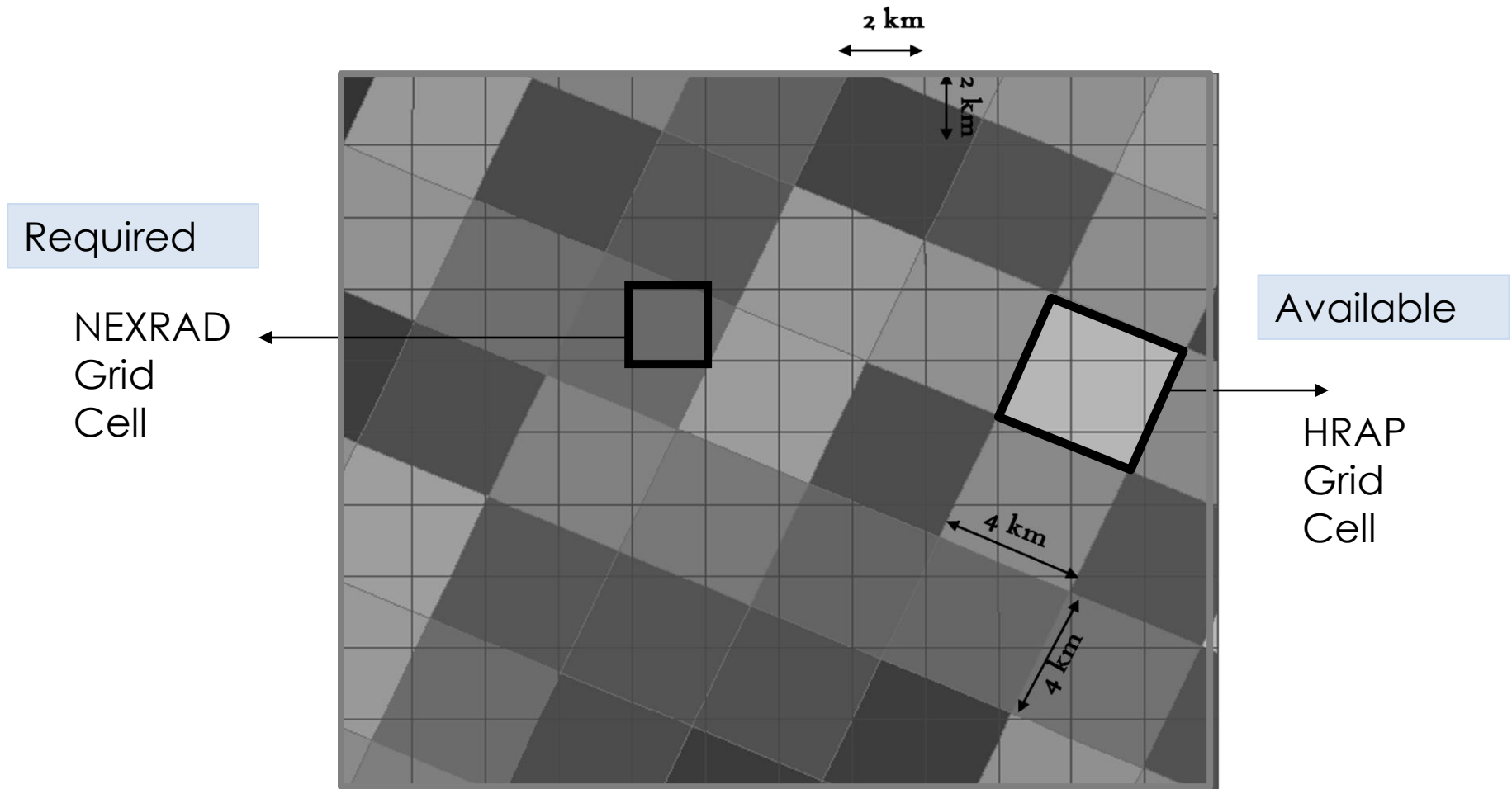
Original and filtered image



Median Filter Exercise

- Use any image
- Add noise
- Filter using mean and median filters
 - Using functions from MATLAB
- Filter demo (available).

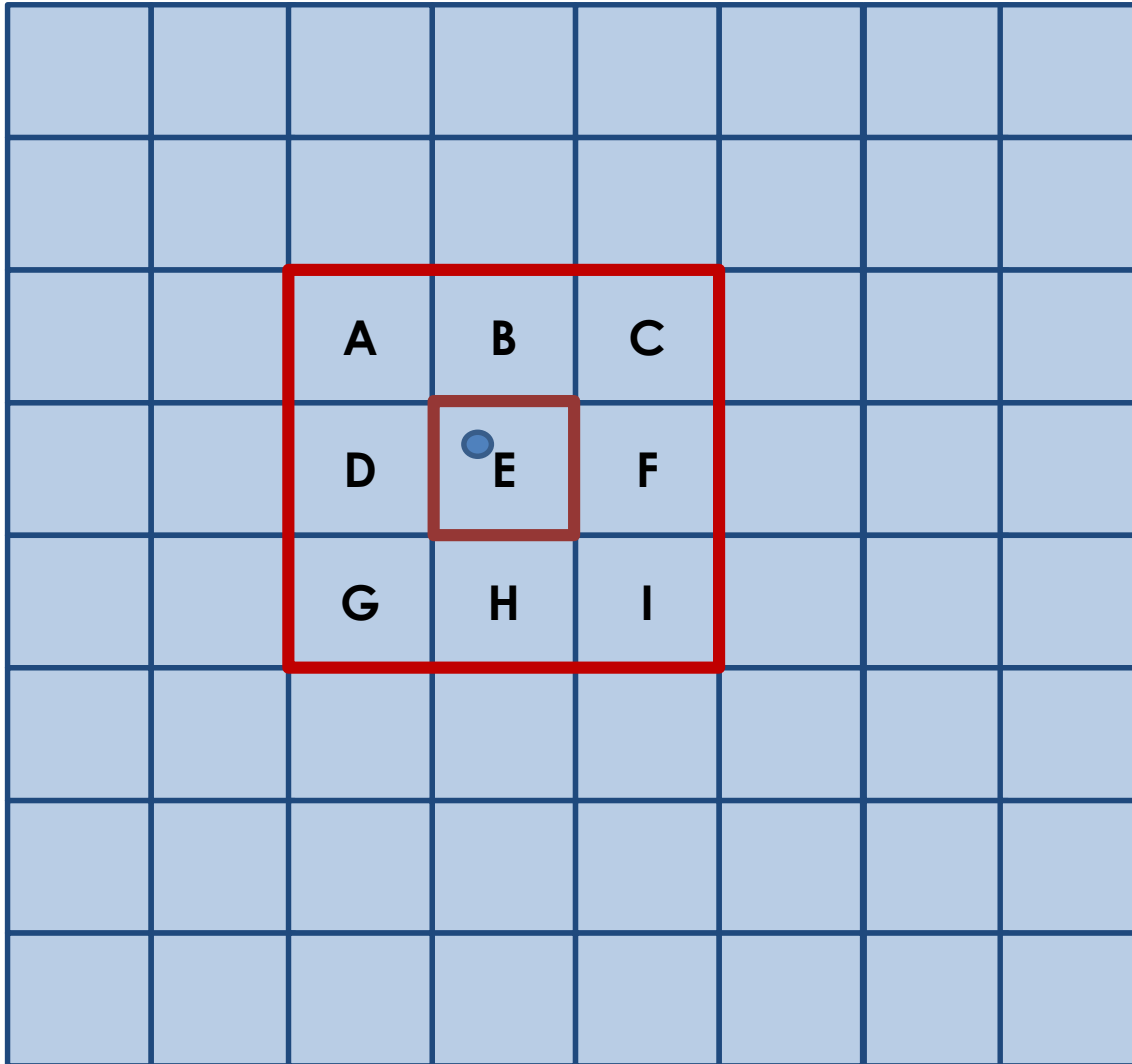
Practical Example



Problem

- Obtain radar estimates for a Cartesian grid of **2 x 2 km** from estimate available at different spatial resolution and orientation (a diagonal **4 x 4 km** grid).
- **Focal operations** ?
- Spatial Interpolation ?

Example of use of local filters: Estimation of missing values



- Estimation of missing values at a rain gauge using radar-based data.
- Develop a method or methods for estimating missing values ?
- How many cells can be used ?
- How can you objectively select the number of cells ?
- Are the cells constant over time and space ?

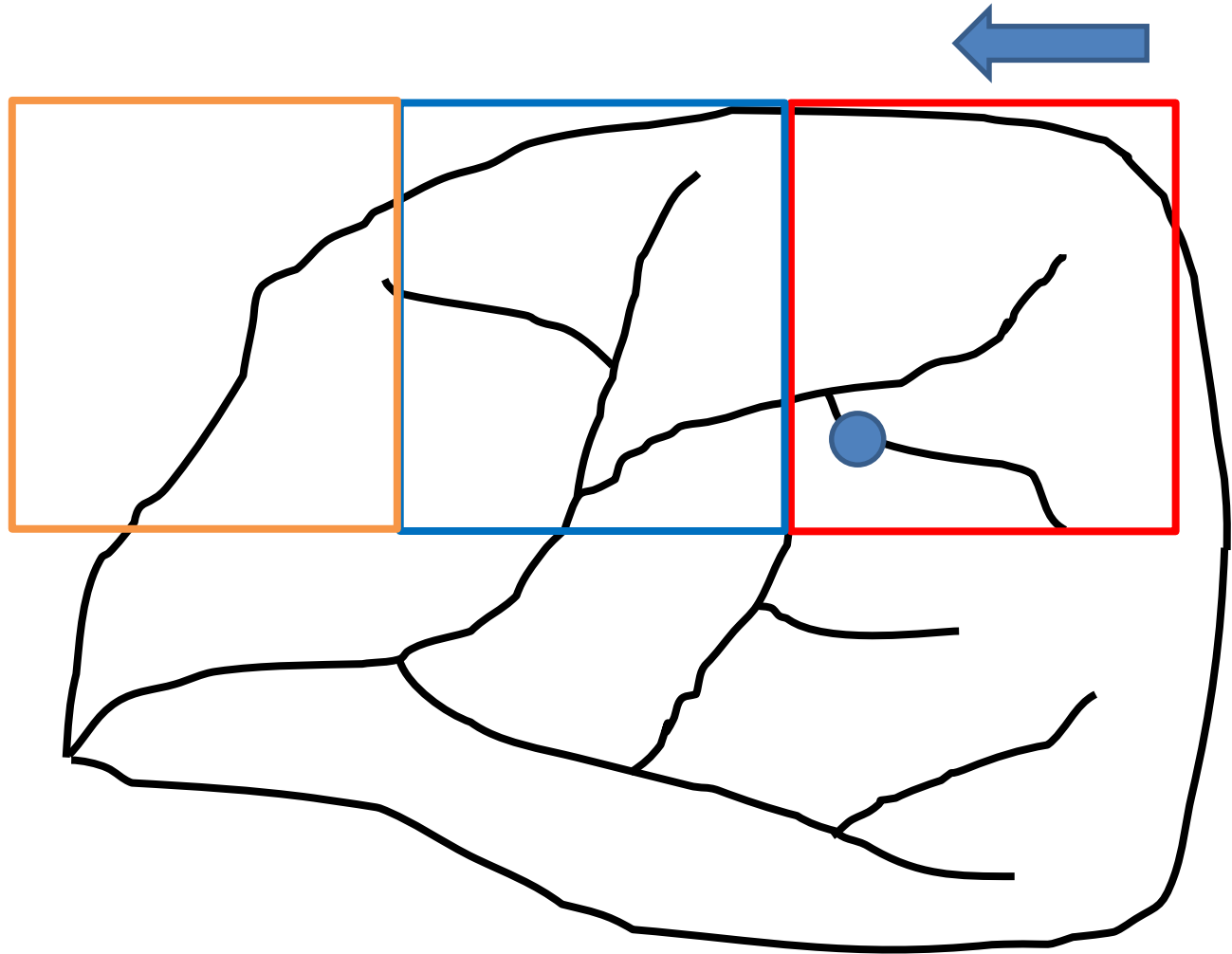
Regression

- Global
- Local
- Geographically Weighted Regression

Local Regression

- As the name suggests regression is based on local spatial attributes to estimate a predictand using a set of predictors.
- Regression is mainly meant for locally varying variables (in a spatial sense).
- The idea is to develop models that are relevant and specific to one single region.

Moving Window Regression (MWR)



Example : Precipitation and
Elevation

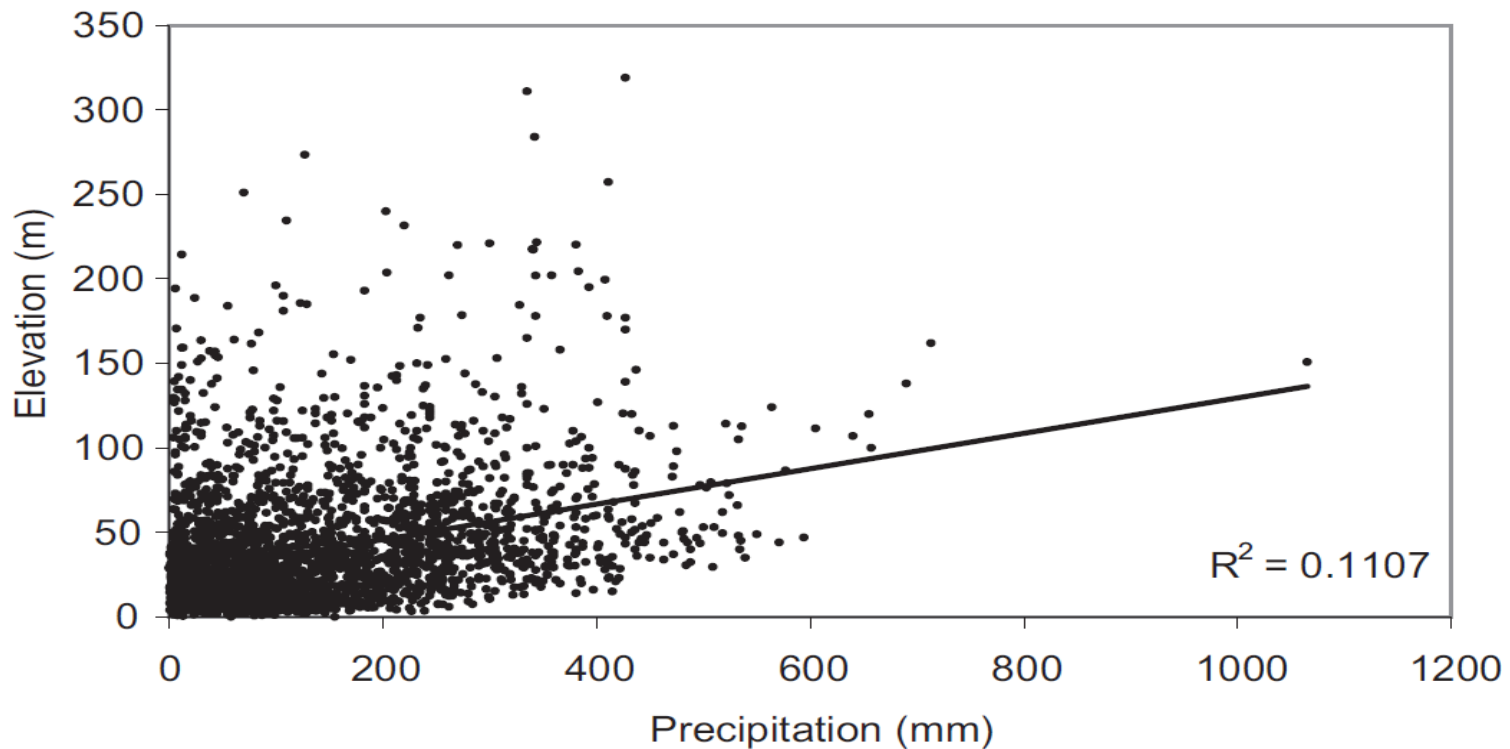
Notes. Copyrighted

(Precipitation-elevation Regressions on Independent Slopes
Model),

MWR (Moving Window Regression)

- The regression is based on the modelled relationship between the data at the several locations closest to the center of the moving window.
- A more sophisticated approach is to weight observations according to their proximity to the center of the window.

Linking Precipitation and Elevation



GWR

- Geographically Weighted Regression⁺
- Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity
- **Spatial nonstationarity** is a condition in which a simple ‘global’ model cannot explain the relationships between some sets of variables.
- The **nature of the model must alter over space** to reflect the structure within the data.
- GWR allows different relationships to exist at different points in space.
- This technique is loosely based on kernel regression.

A simple Linear Regression

$$y_i = a_0 + \sum_{k=1,m} a_k x_{ik} + \varepsilon_i$$

Least squares method is used to obtain the coefficients.

$$\hat{\mathbf{a}} = (\mathbf{x}^t \mathbf{x})^{-1} \mathbf{x}^t \mathbf{y}$$

¹ For the a_0 term a column of 1s must be included in \mathbf{x} .

GWR

- GWR is a relatively simple technique that extends the traditional regression framework of equation (shown in earlier slide) by allowing local variations in rates of change so that the coefficients in the model rather than being global estimates are specific to a location i .

GWR

$$y_i = a_{i0} + \sum_{k=1,m} a_{ik}x_{ik} + \varepsilon_i$$

Note that the parameters are now varying with the location i

$$\tilde{\mathbf{a}} = (\mathbf{x}^t \mathbf{w} \mathbf{x})^{-1} \mathbf{x}^t \mathbf{w} \mathbf{y}.$$

$$\mathbf{a}(\mathbf{i}) = (\mathbf{x}^t \mathbf{w}(\mathbf{i}) \mathbf{x})^{-1} \mathbf{x}^t \mathbf{w}(\mathbf{i}) \mathbf{y}.$$

Weighting approach to specific the availability of observations to point
Of interest i

Weight Assignments

Based on the distance (pre-specified)

$$w_{ij} = 1 \quad \text{if} \quad d_{ij} < d;$$

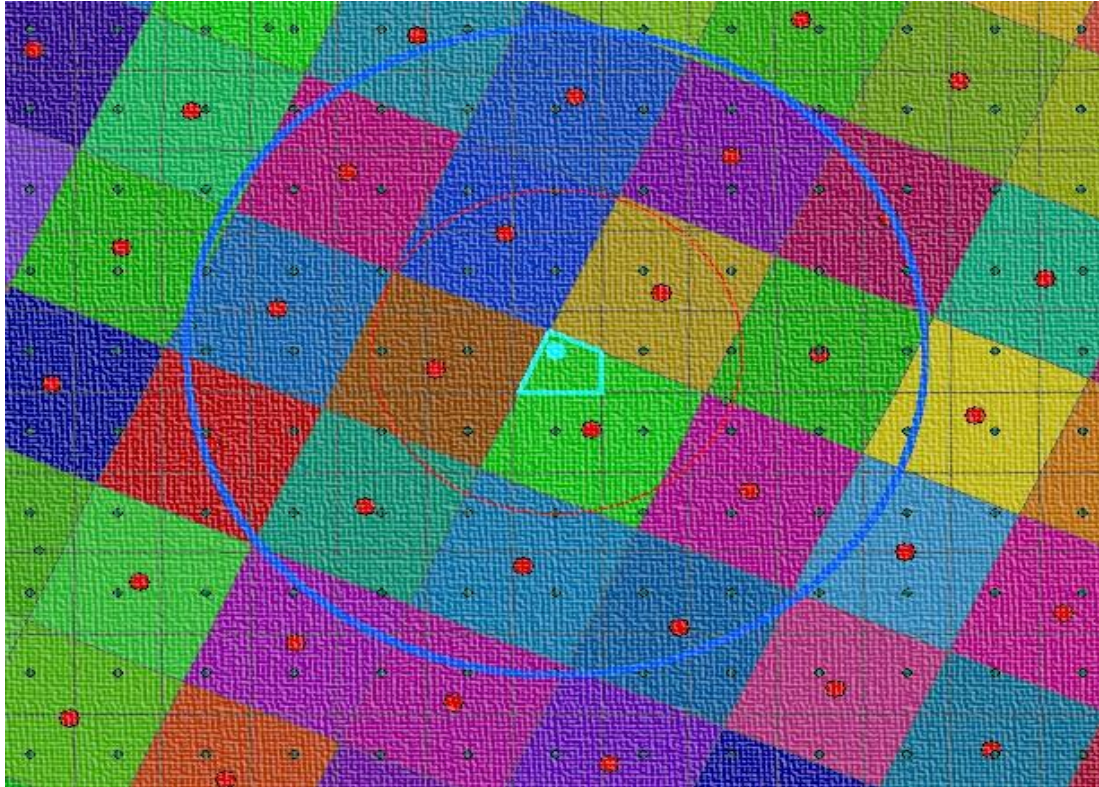
$$w_{ij} = 0 \quad \text{otherwise.}$$

A continuous function can be used.

$$w_{ij} = \exp(-\beta d_{ij}^2)$$

Applications

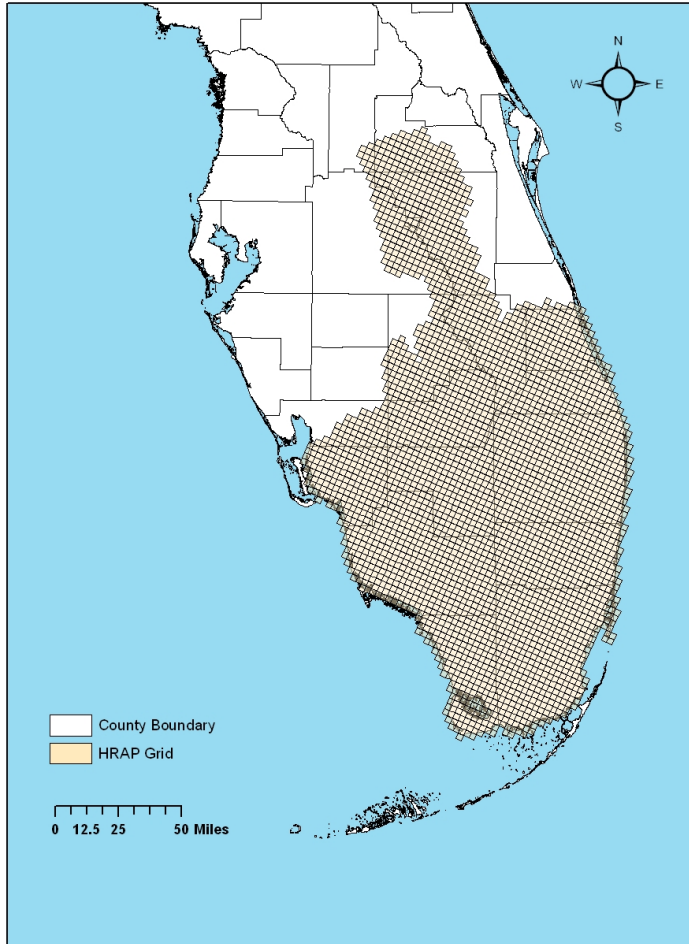
An overlay of HRAP and NEXRAD (2 km x 2 km) grids



GEO-SPATIAL GRID-BASED TRANSFORMATIONS OF MULTI-SENSOR PRECIPITATION ESTIMATES

Objectives

- Evaluation of Method of geometric transformation of HRAP grid to Cartesian grid
- Evaluation of available spatial analysis techniques for achieving the transformation
- Development of multiple spatial transformations/approaches, and test and evaluate the performance of each of these approaches.
- Implementation of the transformation approach to obtain estimates of rainfall for 2 km x 2km Cartesian grid for the period of interest (years 1995 -2001)

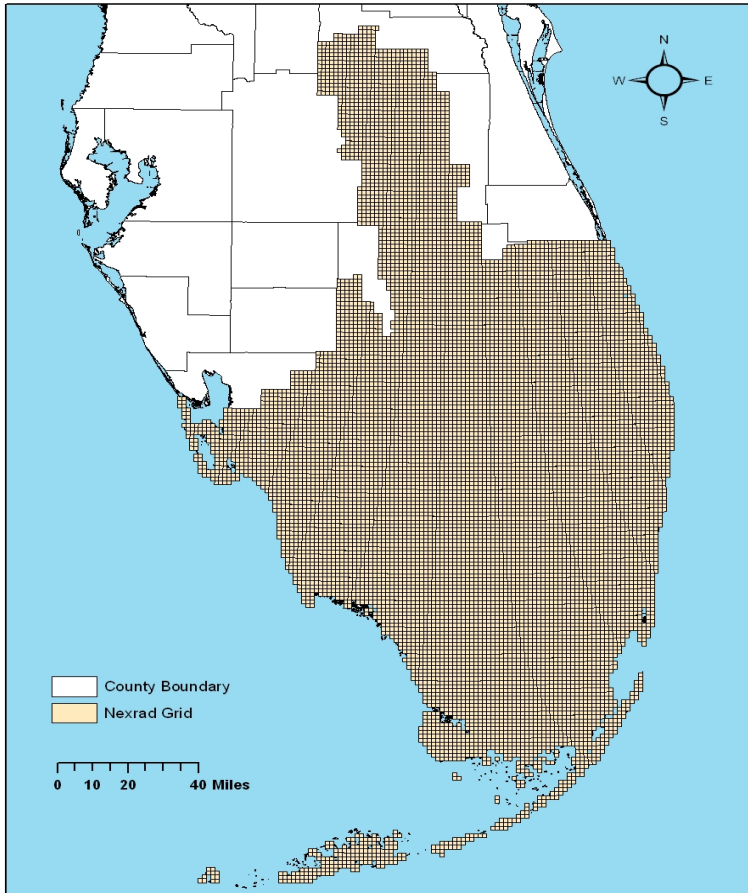


The MPE product created at FSU combines NWS radar and gauge data at hourly.

Rain gauge data were provided by the WMDs (water management districts) and the National Climatic Data Center (NCDC). Although the gauge data had been quality controlled by their respective agencies, FSU conducted a second quality control (QC) described by Marzen and Fuelberg (2005).

HRAP (4km x 4km) Product

The radar data were provided by the NWS Southeast River Forecast Center (SERFC). The final hourly MPE product, called MMOSAIC, is placed on the Hydrologic Rainfall Analysis Project (HRAP) ~ 4×4 km grid.



The OneRain Corporation provided its rainfall database to the five Water Management Districts in Florida and to FSU.

The final OneRain product is placed on a 2×2 km Cartesian grid and the data are provided at 15 min. intervals.

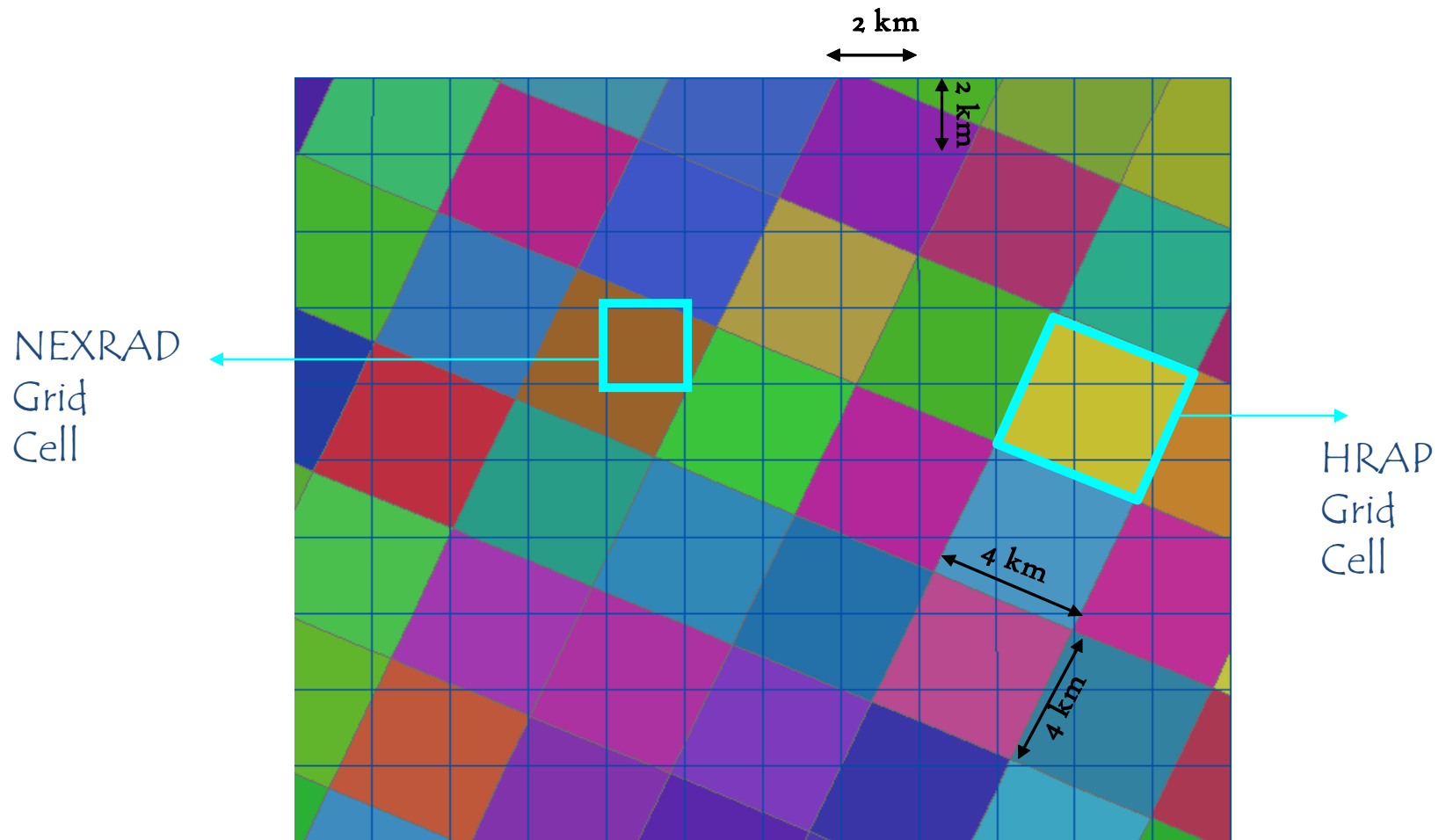
NEXRAD Data (2km x 2km)

The OneRain Corporation provides a near real-time product as well as an end-of-the-month product which has undergone further QC. The OneRain Corporation has provided the SFWMD with NEXRAD based precipitation estimates from years 2002 to 2007 (Pathak, 2001). However, data on a 2 km x 2km Cartesian grid before year 2001 was not available to SFWMD.

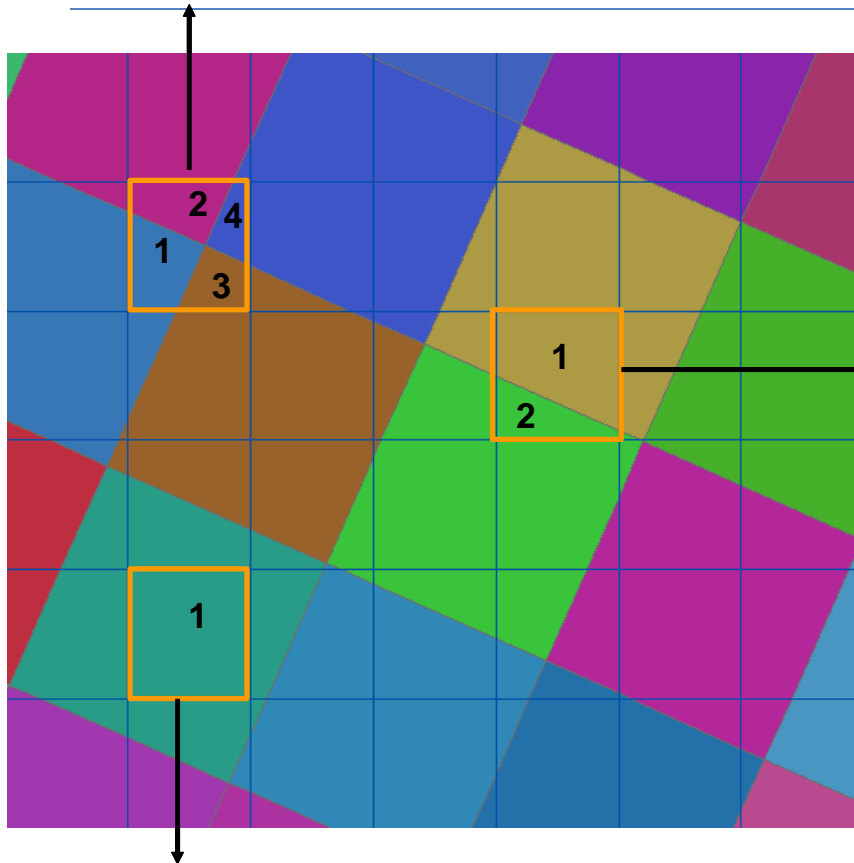
Interpolation Methods

- Area Weighting Method
- Maximum Area Method
- Inverse Distance Weighting Method
- Fixed Radius Distance Weighting Method
- Inverse Exponential Weighting Method
- Equal Weights (Average) Method
- Kriging?
- Other Methods?

Overlay of NEXRAD Grid over HRAP grid



4 HRAP cells (areas) in 1 NEXRAD Grid



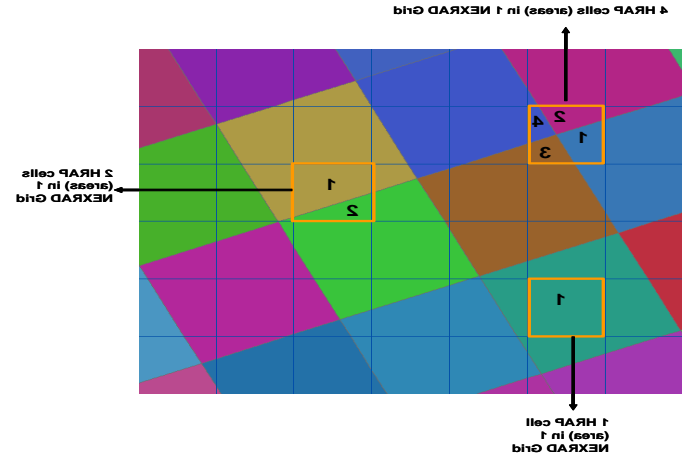
2 HRAP cells
(areas) in 1
NEXRAD Grid

1 HRAP cell
(area) in 1
NEXRAD Grid

Area based Weighting

In this method, the aerial extent of overlay is used as a weight.

$$\theta_i = \frac{\sum_{j=1}^n A_j \phi_j}{\sum_{j=1}^n A_j} \quad \forall i$$



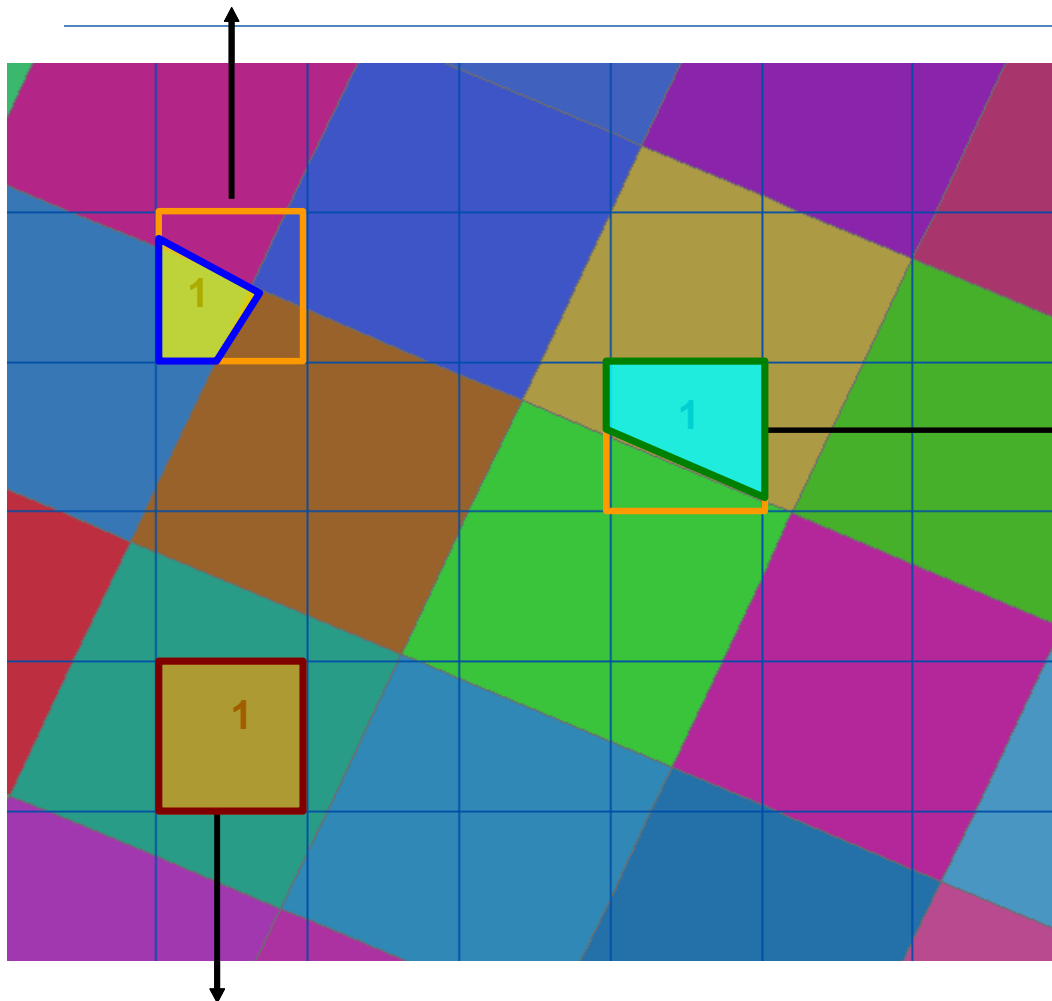
where i = index for NEXRAD(2km x 2km) Cartesian grid

j = index for 4km X 4km HRAP grid

A_j = Area of 4km X 4km HRAP grid, j , within a 2km x 2km NEXRAD grid, i .

n = number of distinct areas of HRAP grid within a 2km x 2km NEXRAD grid, i .

Maximum HRAP cell (area) in 1 NEXRAD Grid



Maximum HRAP
cell (area) in 1
NEXRAD Grid

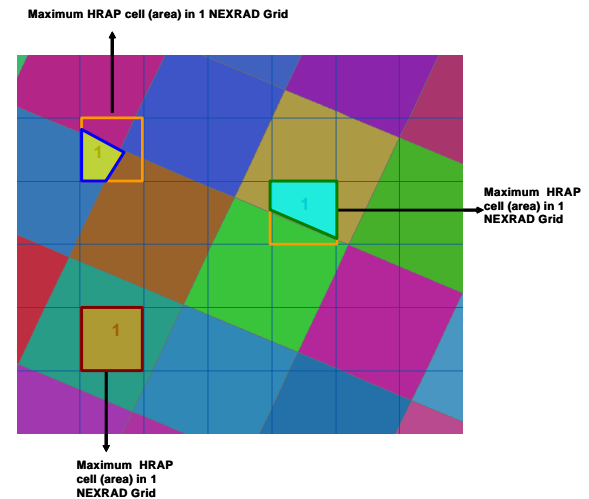
Maximum HRAP
cell (area) in 1
NEXRAD Grid

Maximum Area

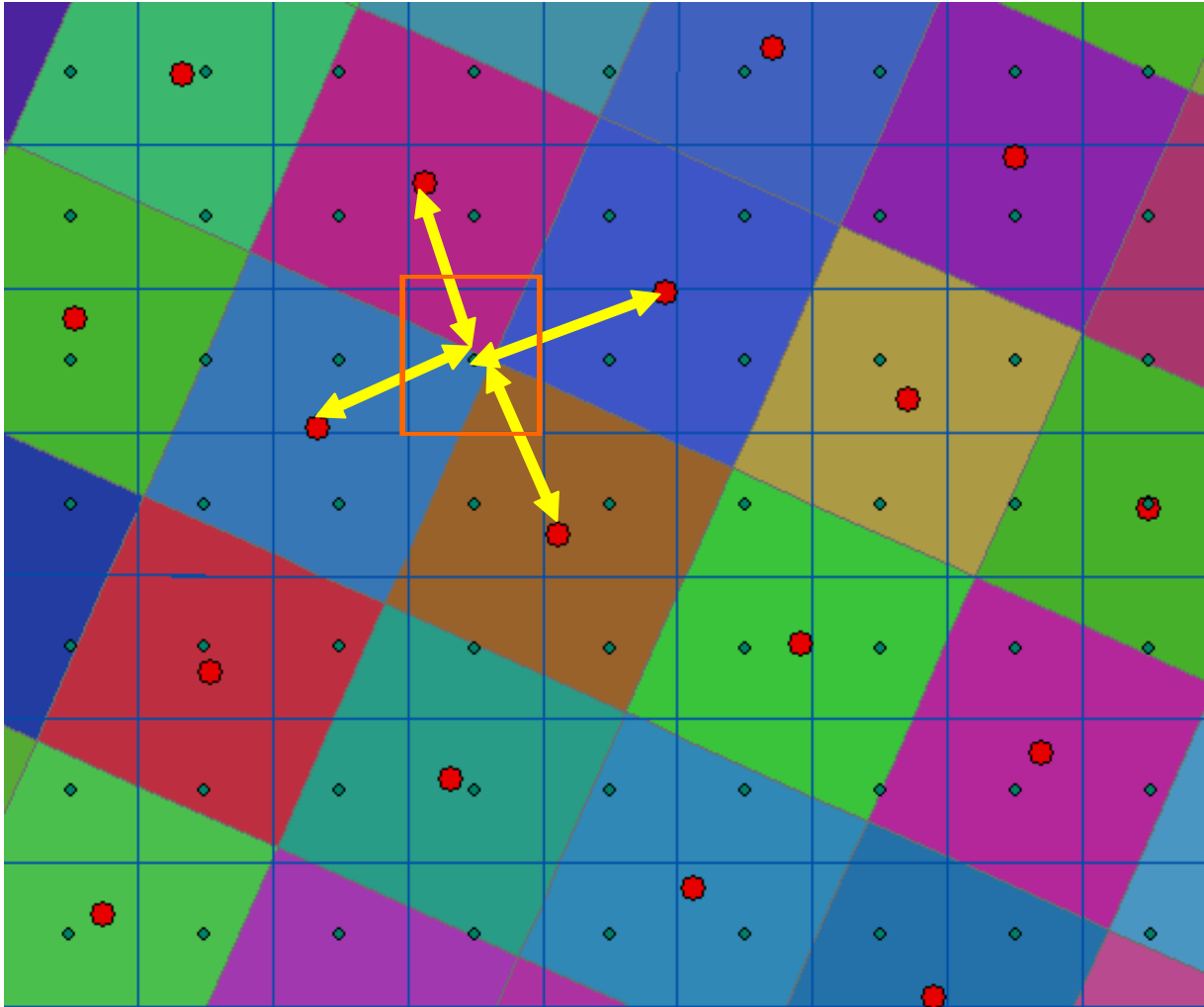
$$\theta_i = \phi_j \quad \forall i$$

$$A_j = \max (A_k) \quad \forall k$$

$$j = k$$

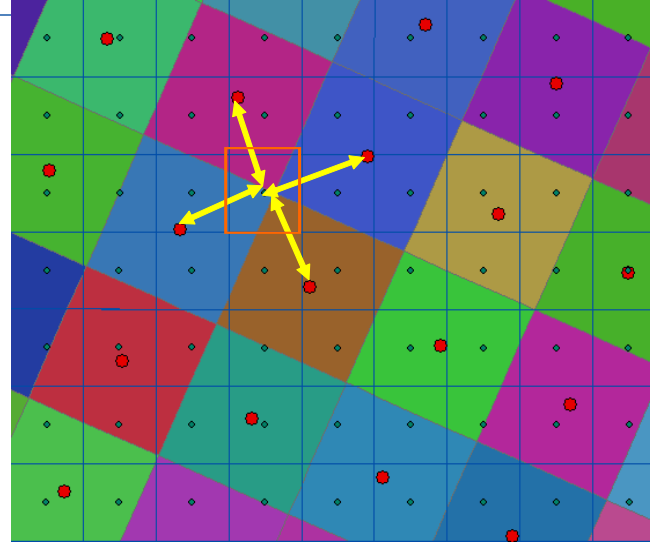


where A_j = maximum Area occupied by HRAP grid within 2km X 2km grid.

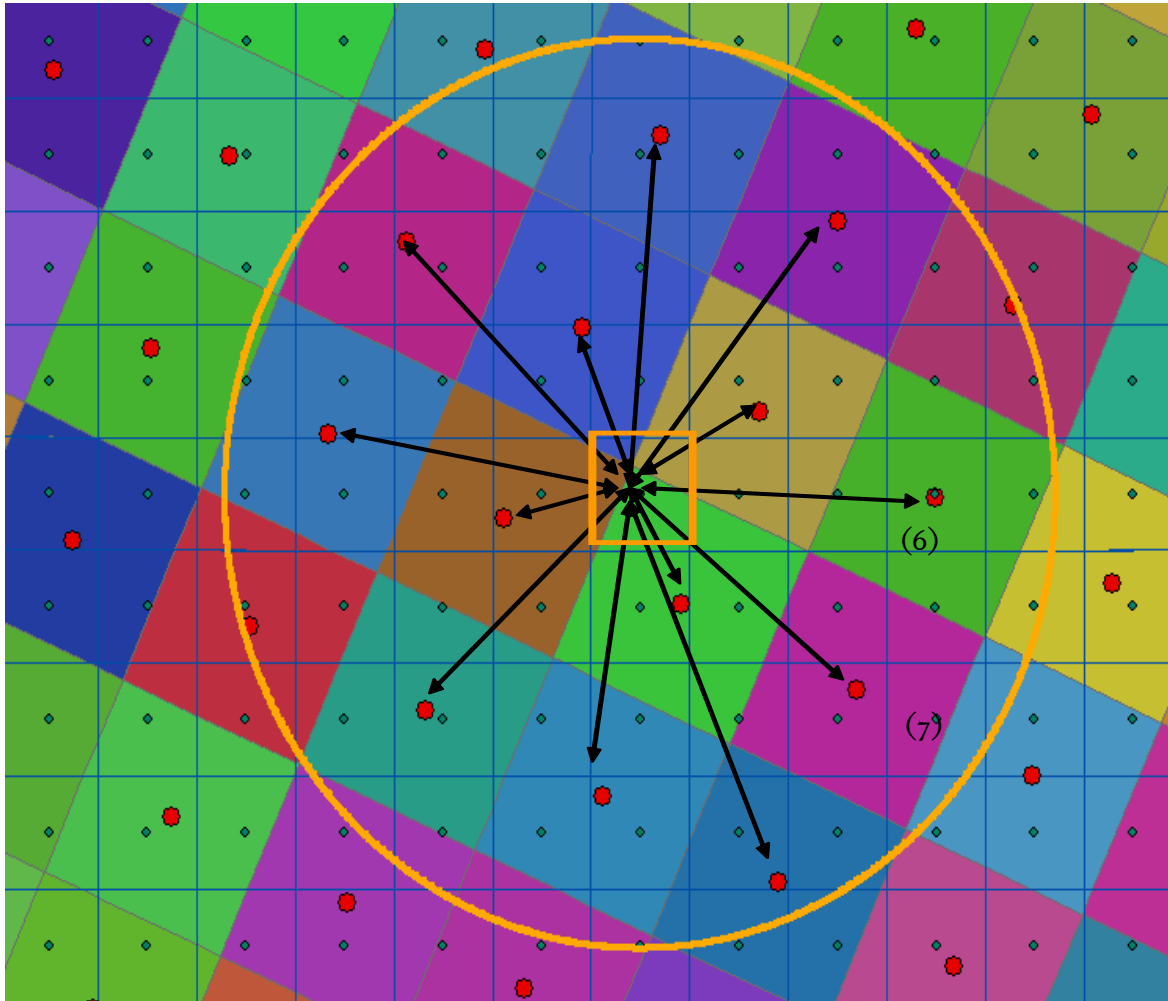


Distance based
~ Inverse Distance

$$\theta_m = \frac{\sum_{i=1}^n \theta_i d_{mi}^{-k}}{\sum_{i=1}^n d_{mi}^{-k}}$$



where again θ_m is the estimate of the observation at a point in space, m ; n is the number of observation points (centroids of HRAP cells); θ_i is the observation of each HRAP cell i , d_{mi} is the distance from the location of centroid of the HRAP cell i to the observation point m ; and k is referred to as friction distance (Vieux, 2001) that ranges from 1.0 to 6.0. In the current study a value of 2 is used for the friction distance, k . In this method distances can be calculated only for those HRAP cells which intersect the NEXRAD cells.



Fixed Radius - IDWM

$$\theta_m = \frac{\sum_{i=1}^n \theta_i d_{mi}^{-k}}{\sum_{i=1}^n d_{mi}^{-k}}$$

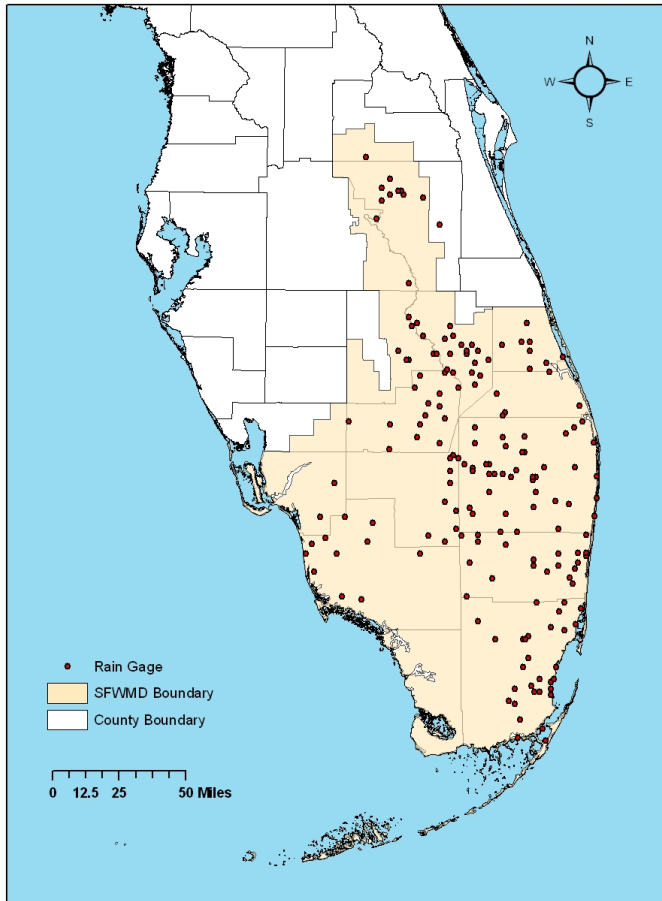
$$d_{mi} \leq D_{fr} \quad \forall i$$

Exponential & Average Weighting Methods

$$\theta_m = \frac{\sum_{i=1}^n \theta_i e^{-kd_{mi}}}{\sum_{i=1}^n e^{-kd_{mi}}}$$

Popular in Geosciences/Geography

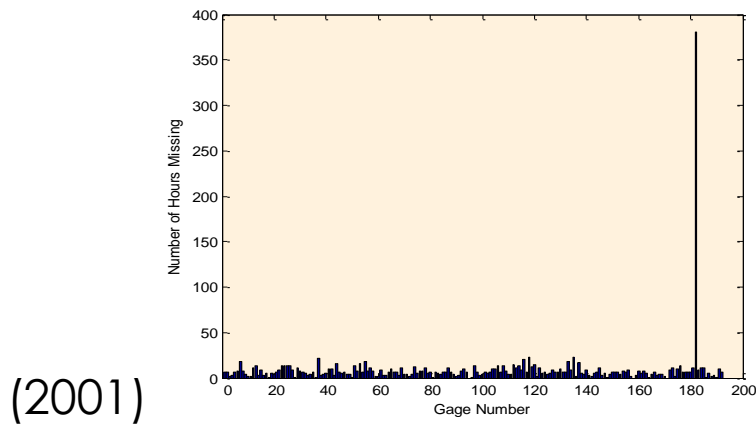
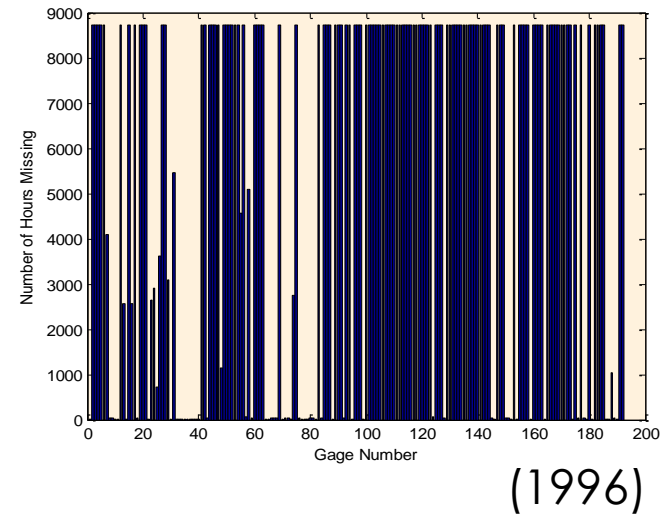
$$\theta_m = \frac{\sum_{i=1}^n \theta_i w_i}{\sum_{i=1}^n w_i}$$



Observations from 192 rain gages available in the South Florida Water Management District (SFWMD) region were used for assessment of the performance of the six methods.

Missing Data

Year	HRAP	Rain Gage
1996		
1997		
1998		
1999	December	January
2000	May 27 - May 31	
2001	April & December	



Error Measures (Performance Measures)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\phi}_i - \phi_i)^2}$$

$$MRE = \frac{1}{n} \sum_{i=1}^n \frac{|\hat{\phi}_i - \phi_i|}{\phi_i}$$

$$AE = \sum_{i=1}^n |\hat{\phi}_i - \phi_i|$$

$$\rho = \frac{\sum (\hat{\phi}_i - \mu_g)(\phi_i - \mu_n)}{(n-1)\sigma_g \sigma_n}$$

Assessment of Error Measures

- The error measures are used to select the best method out of the six methods discussed. The use of several error measures provides several advantages as well few disadvantages in the method selection process.
- The advantages include 1) assessment of performance of methods using different indices, 2) evaluation of error structure and correlation between observed and estimated.
- The main disadvantage is that there is no absolute way of selecting the best method.

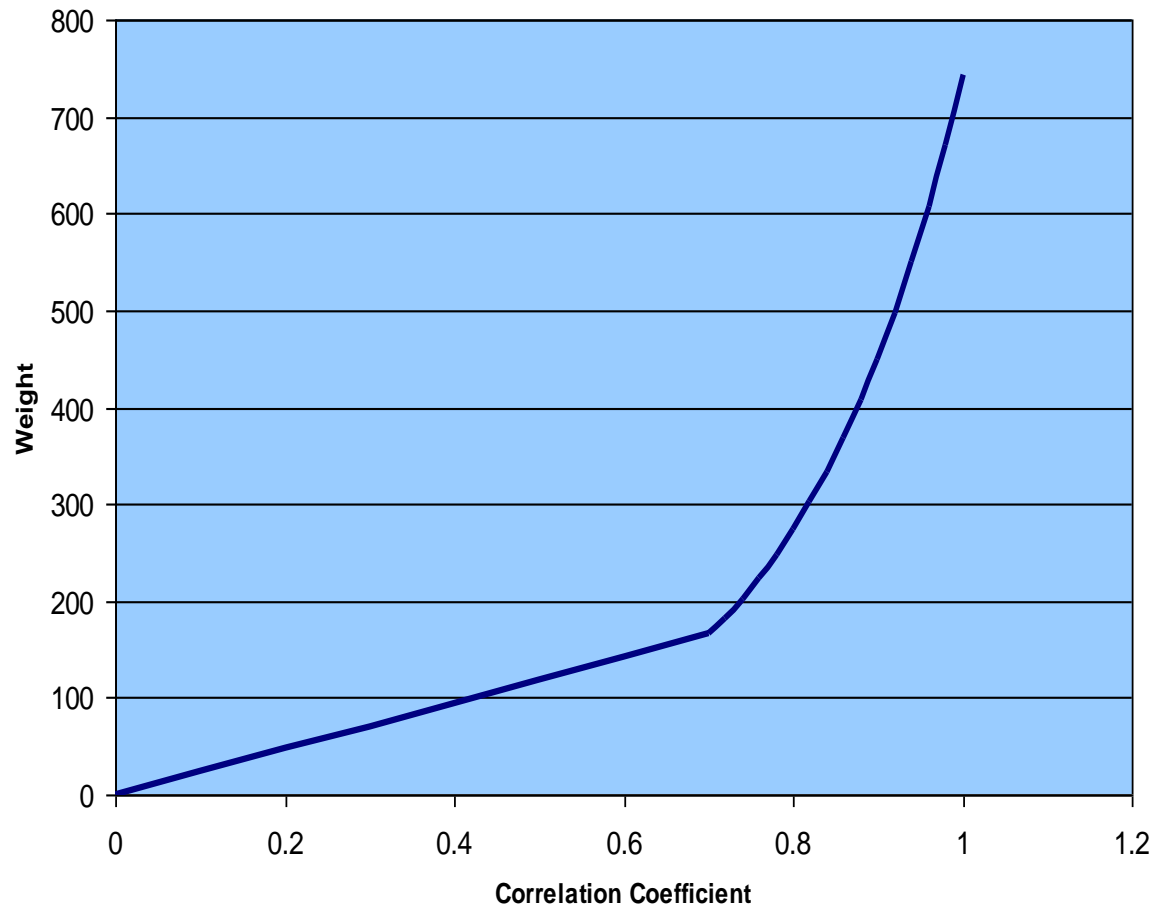
Weighting Function (Membership Functions)

- A method by which the error measures are transformed to a common dimensionless parameter that can be used for selection process is required.
- In the current study weighting functions are proposed as a way of generating non-dimensional weights for each of the error/performance measures.
- Similar measures using fuzzy set theory were developed by Teegavarapu et. al, (2006, Journal of Hydroinformatics)

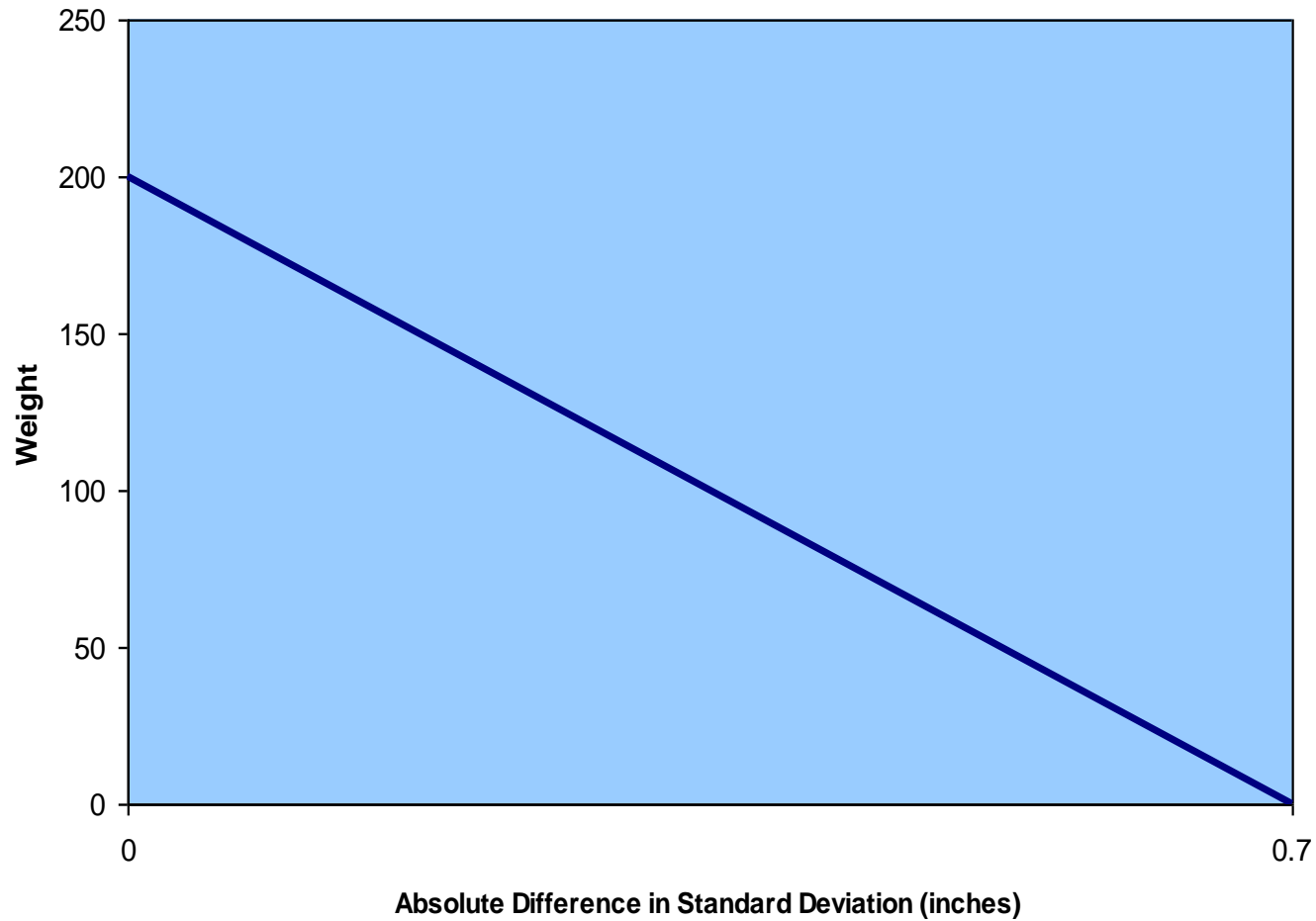
Weighting Functions

- The functions are designed in such as way, that the maximum value is always attached to the best performance based on a specific error measure. Linear and non-linear weighting functions are developed considering the upper and lower bounds of each performance measure. The functions are chosen carefully considering the importance attached to each of the error measures

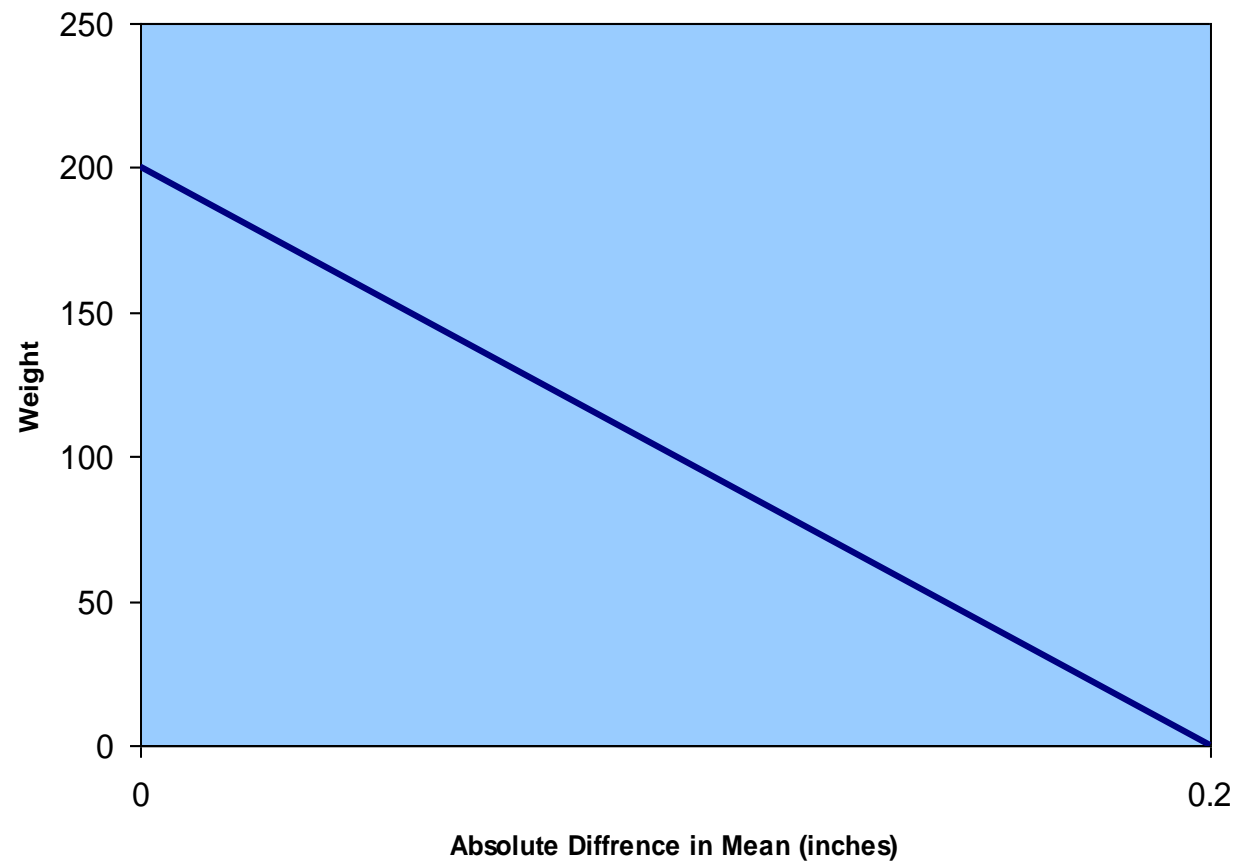
Correlation Coefficient



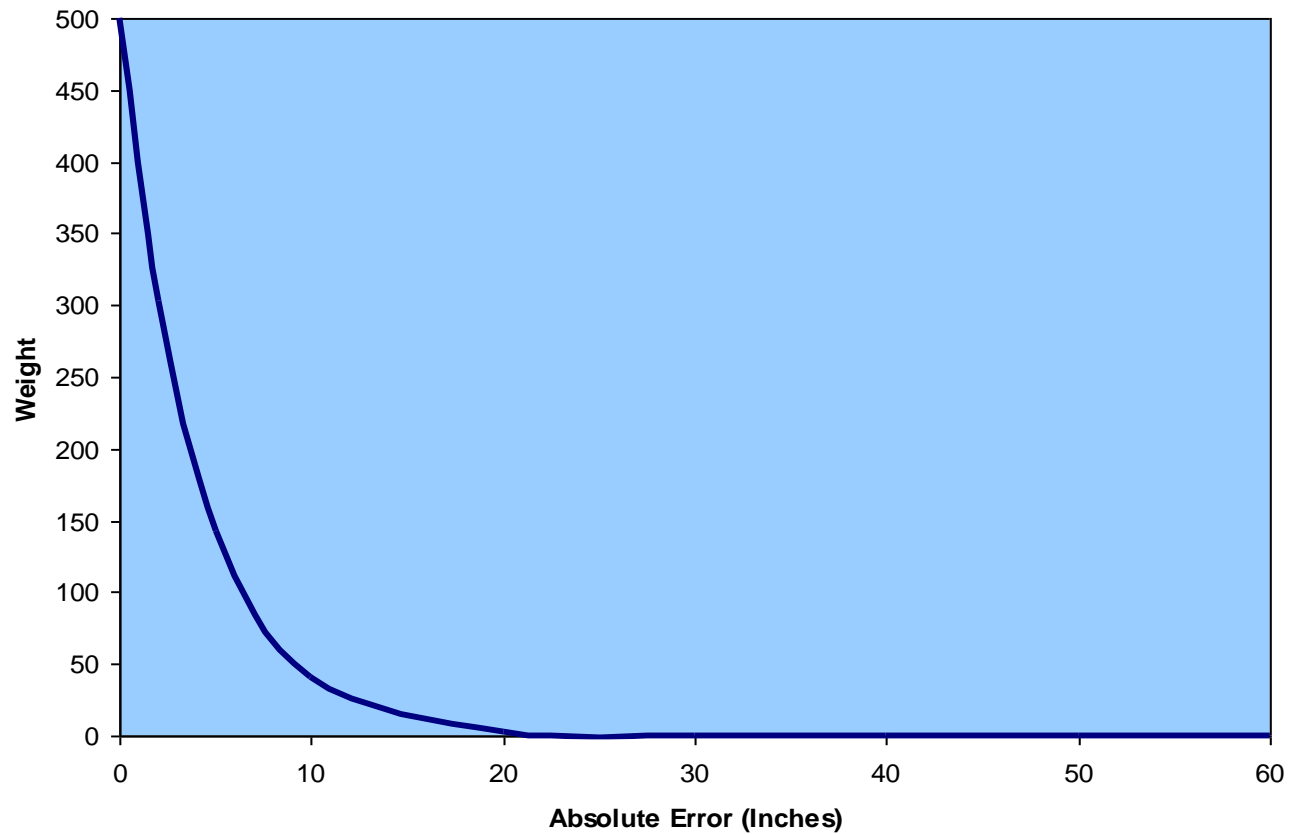
Difference in Standard Deviation

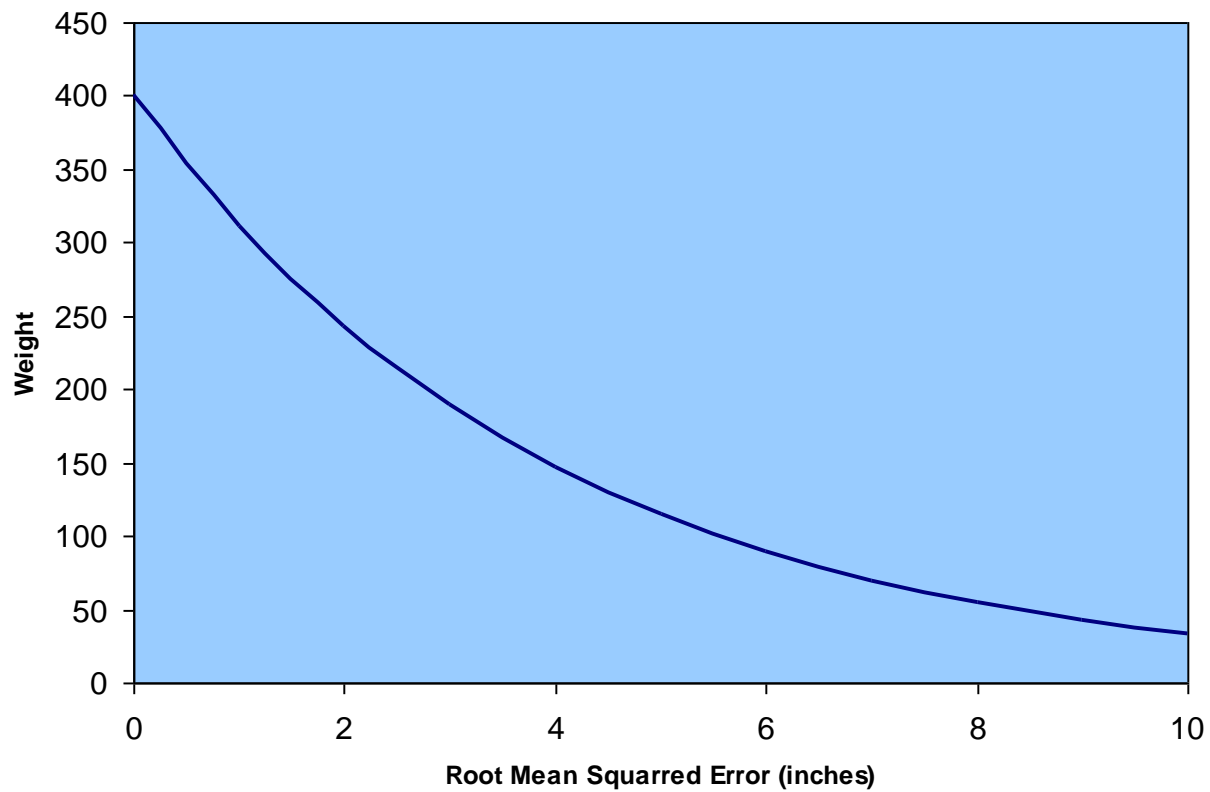


Difference in Mean



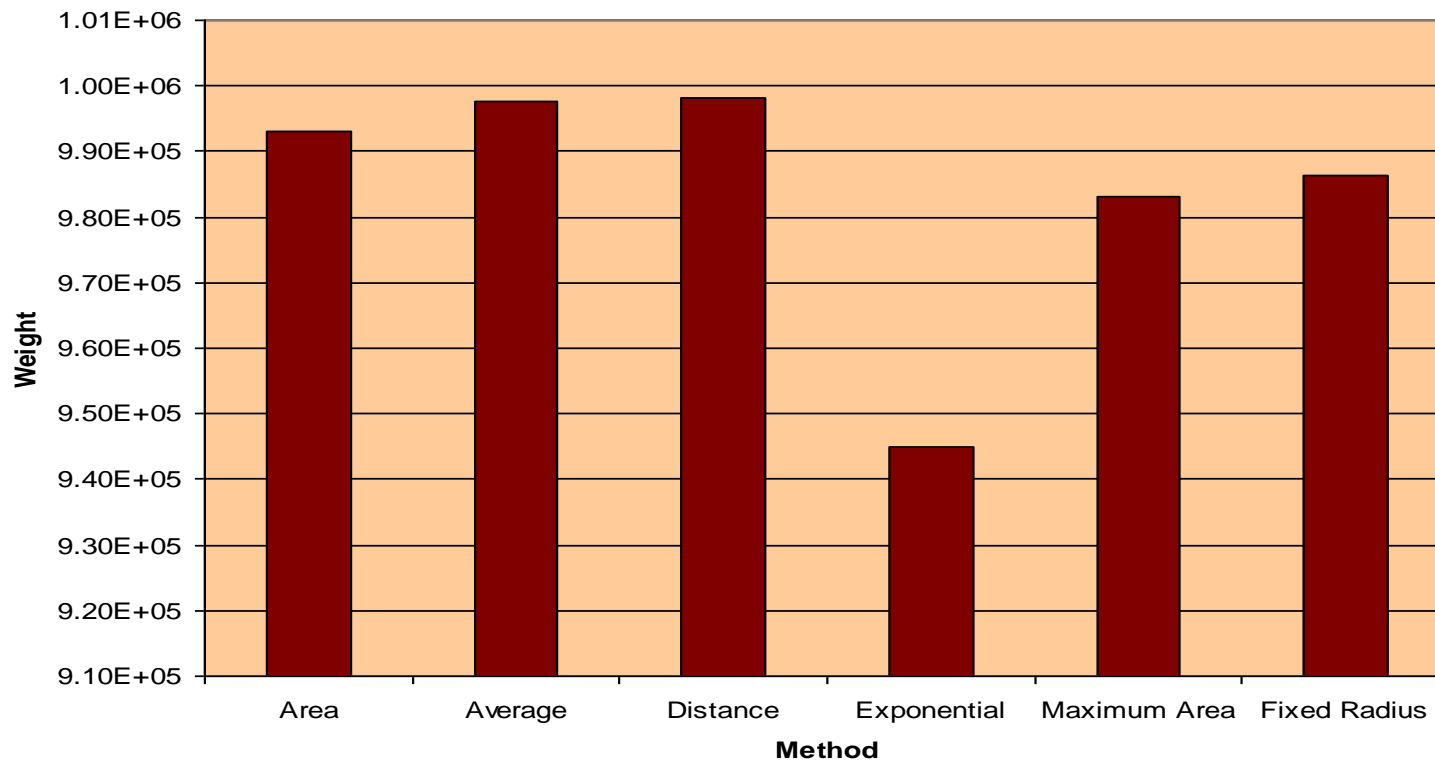
Absolute Error

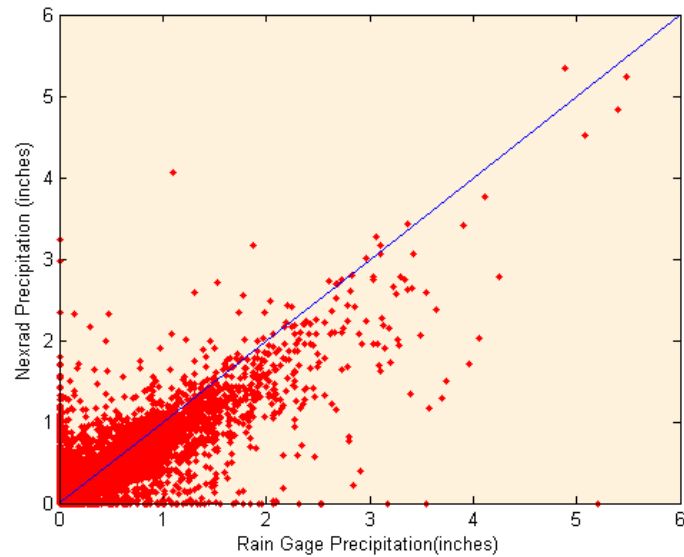




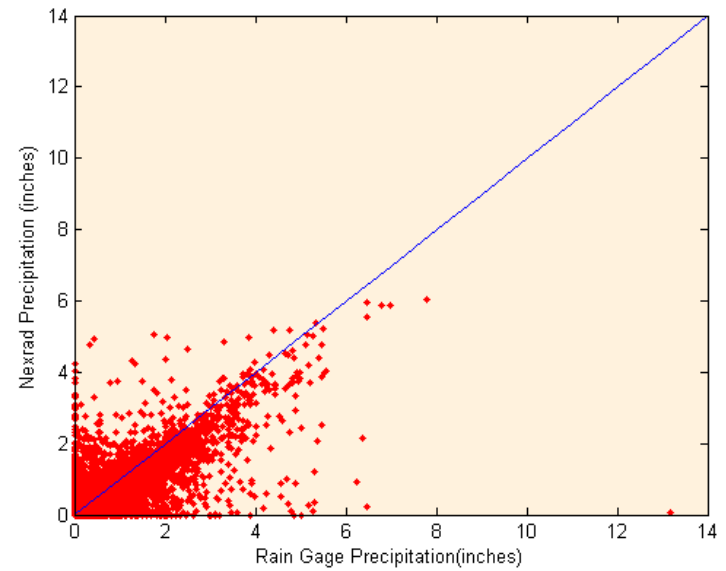
Year	Method	Weight							Total
		W_u	W_g	W_{cv}	W_D	$W_{e'}$	W_{RMSE}	W_{MRE}	
1996	Area	1.49E+04	1.47E+04	5.23E+03	2.58E+04	6.05E+02	7.67E+03	1.62E+04	8.51E+04
	Average	1.49E+04	1.47E+04	5.02E+03	2.66E+04	5.49E+02	7.79E+03	1.63E+04	8.58E+04
	Distance	1.49E+04	1.47E+04	5.15E+03	2.63E+04	5.97E+02	7.78E+03	1.63E+04	8.57E+04
	Exponential	1.49E+04	1.46E+04	3.98E+03	2.44E+04	2.66E+02	6.84E+03	1.57E+04	8.07E+04
	Maximum Area	1.49E+04	1.47E+04	5.49E+03	2.48E+04	6.15E+02	7.50E+03	1.60E+04	8.40E+04
	Fixed Radius	1.49E+04	1.47E+04	4.70E+03	2.62E+04	4.25E+02	7.46E+03	1.62E+04	8.46E+04
1997	Area	3.18E+04	3.13E+04	1.21E+04	4.72E+04	1.20E+03	1.10E+04	3.05E+04	1.65E+05
	Average	3.18E+04	3.13E+04	1.17E+04	4.82E+04	1.06E+03	1.11E+04	3.06E+04	1.66E+05
	Distance	3.18E+04	3.13E+04	1.19E+04	4.80E+04	1.16E+03	1.12E+04	3.06E+04	1.66E+05
	Exponential	3.19E+04	3.12E+04	9.71E+03	4.42E+04	8.21E+02	9.34E+03	2.96E+04	1.57E+05
	Maximum Area	3.18E+04	3.13E+04	1.25E+04	4.55E+04	1.27E+03	1.10E+04	3.02E+04	1.64E+05
	Fixed Radius	3.19E+04	3.13E+04	1.11E+04	4.77E+04	1.00E+03	1.06E+04	3.05E+04	1.64E+05
1998	Area	3.39E+04	3.34E+04	1.15E+04	5.10E+04	1.32E+03	1.31E+04	3.16E+04	1.76E+05
	Average	3.39E+04	3.33E+04	1.12E+04	5.20E+04	1.18E+03	1.33E+04	3.17E+04	1.77E+05
	Distance	3.39E+04	3.34E+04	1.14E+04	5.17E+04	1.28E+03	1.33E+04	3.17E+04	1.77E+05
	Exponential	3.39E+04	3.32E+04	9.28E+03	4.91E+04	9.96E+02	1.15E+04	3.08E+04	1.69E+05
	Maximum Area	3.39E+04	3.34E+04	1.20E+04	4.97E+04	1.42E+03	1.30E+04	3.13E+04	1.75E+05
	Fixed Radius	3.39E+04	3.33E+04	1.05E+04	5.18E+04	1.16E+03	1.27E+04	3.17E+04	1.75E+05
1999	Area	3.54E+04	3.48E+04	1.21E+04	4.83E+04	3.97E+02	1.13E+04	3.26E+04	1.75E+05
	Average	3.54E+04	3.47E+04	1.16E+04	4.93E+04	2.87E+02	1.14E+04	3.28E+04	1.75E+05
	Distance	3.54E+04	3.47E+04	1.19E+04	4.92E+04	3.64E+02	1.14E+04	3.28E+04	1.76E+05
	Exponential	3.54E+04	3.46E+04	9.08E+03	4.58E+04	1.23E+02	9.54E+03	3.19E+04	1.66E+05
	Maximum Area	3.53E+04	3.48E+04	1.27E+04	4.64E+04	4.72E+02	1.12E+04	3.24E+04	1.73E+05
	Fixed Radius	3.54E+04	3.47E+04	1.08E+04	4.90E+04	2.56E+02	1.08E+04	3.27E+04	1.74E+05
2000	Area	3.79E+04	3.72E+04	1.12E+04	5.18E+04	1.62E+03	1.80E+04	3.80E+04	1.96E+05
	Average	3.79E+04	3.72E+04	1.05E+04	5.34E+04	1.55E+03	1.81E+04	3.83E+04	1.97E+05
	Distance	3.79E+04	3.72E+04	1.09E+04	5.28E+04	1.61E+03	1.81E+04	3.82E+04	1.97E+05
	Exponential	3.80E+04	3.71E+04	7.08E+03	4.89E+04	1.08E+03	1.68E+04	3.73E+04	1.86E+05
	Maximum Area	3.79E+04	3.72E+04	1.19E+04	4.96E+04	1.72E+03	1.80E+04	3.76E+04	1.94E+05
	Fixed Radius	3.80E+04	3.72E+04	9.41E+03	5.23E+04	1.38E+03	1.77E+04	3.81E+04	1.94E+05
2001	Area	3.80E+04	3.73E+04	1.28E+04	5.74E+04	5.22E+02	1.30E+04	3.72E+04	1.96E+05
	Average	3.80E+04	3.73E+04	1.22E+04	5.86E+04	3.87E+02	1.32E+04	3.73E+04	1.97E+05
	Distance	3.80E+04	3.73E+04	1.26E+04	5.83E+04	4.87E+02	1.32E+04	3.74E+04	1.97E+05
	Exponential	3.80E+04	3.71E+04	9.61E+03	5.39E+04	1.46E+02	1.13E+04	3.61E+04	1.86E+05
	Maximum Area	3.79E+04	3.73E+04	1.34E+04	5.50E+04	6.01E+02	1.26E+04	3.68E+04	1.94E+05
	Fixed Radius	3.80E+04	3.73E+04	1.14E+04	5.80E+04	3.49E+02	1.26E+04	3.72E+04	1.95E+05

Evaluation of Different Methods



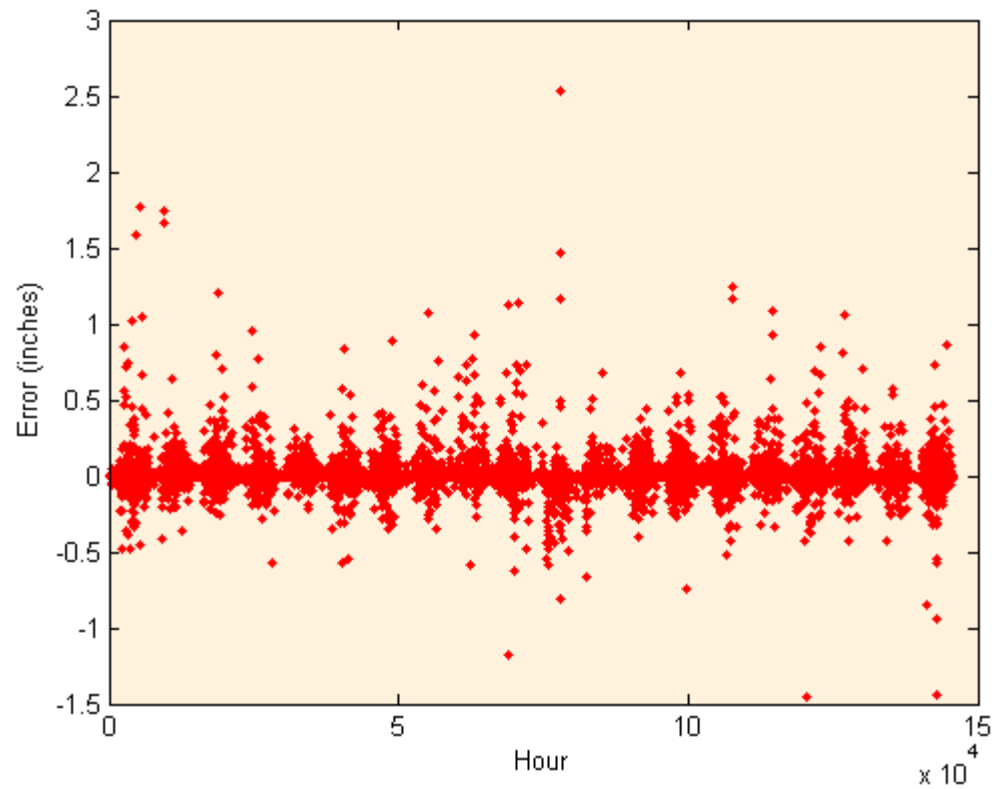


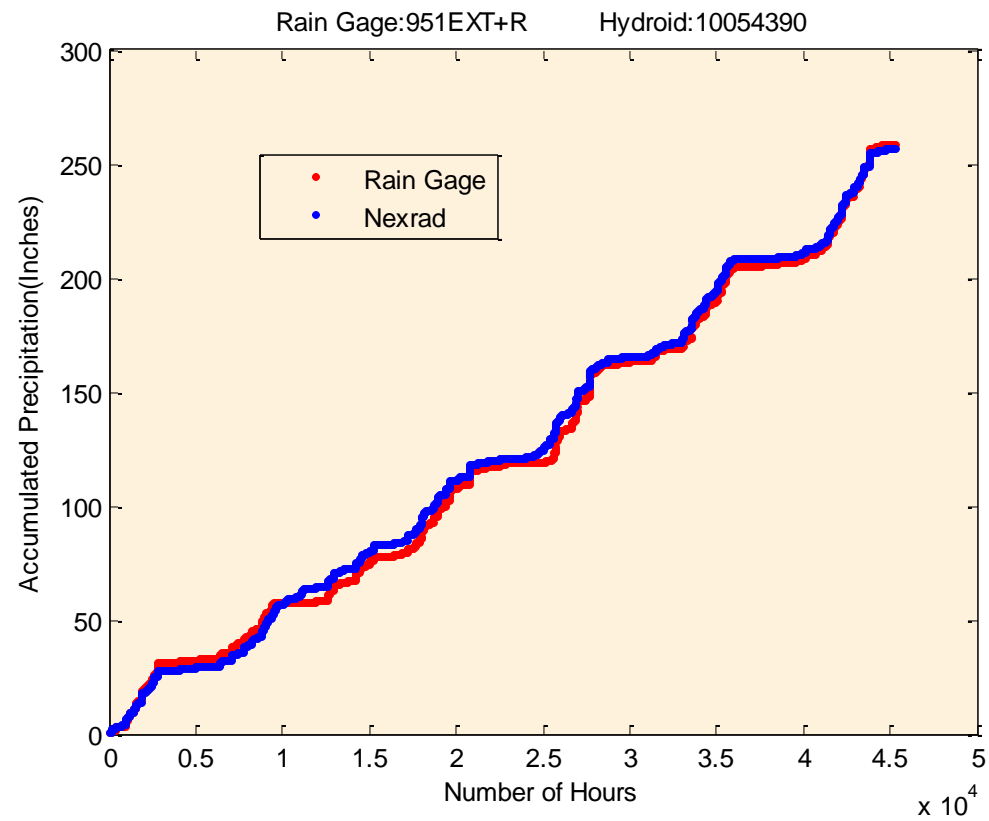
(1996)

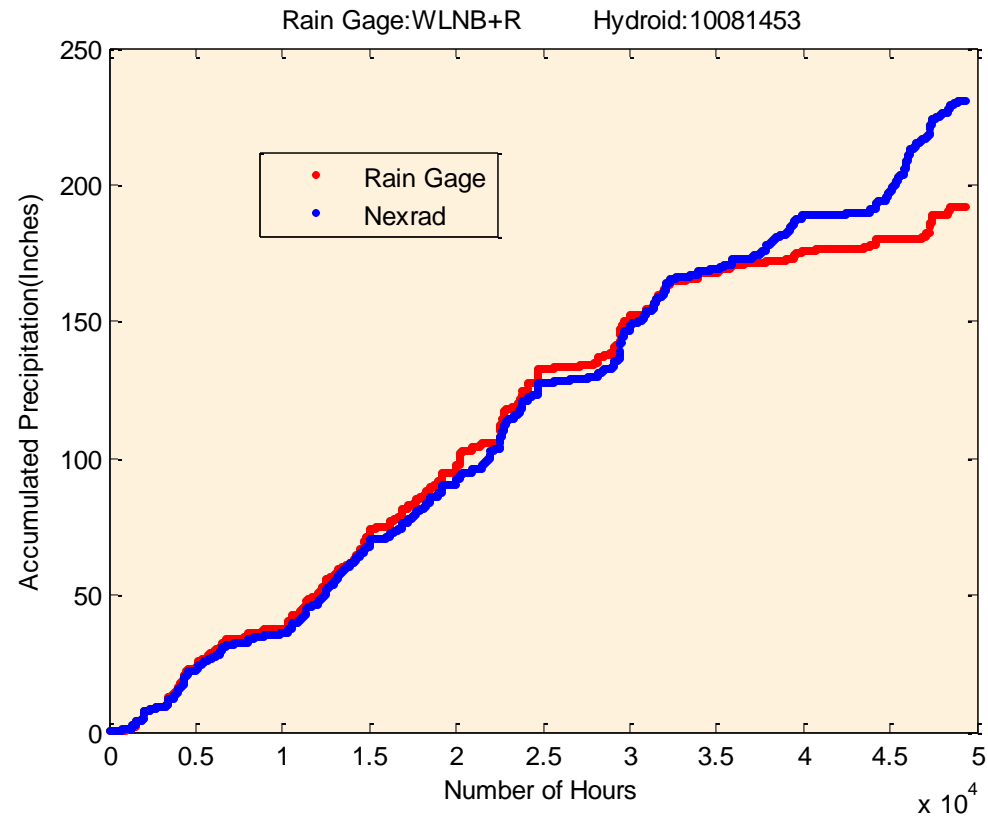


(2001)

Residuals







Bias Corrections

$$\varphi = \frac{\sum_{i=1}^n \frac{G_i}{R_i}}{n} \quad \forall i \quad \beta = \frac{\sum_{i=1}^n G_i}{\sum_{i=1}^n R_i} \quad \forall i$$

$$\theta_j'' = \theta_j * \varphi$$

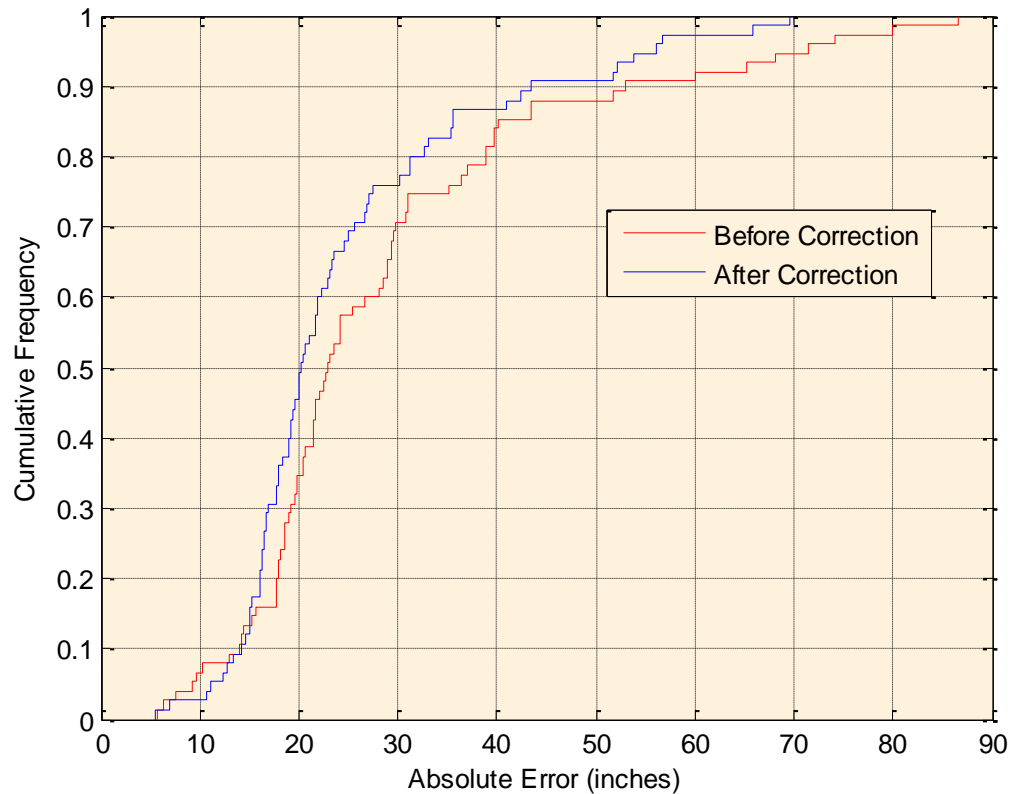
G_i = gage value, R_i = radar value in pixel

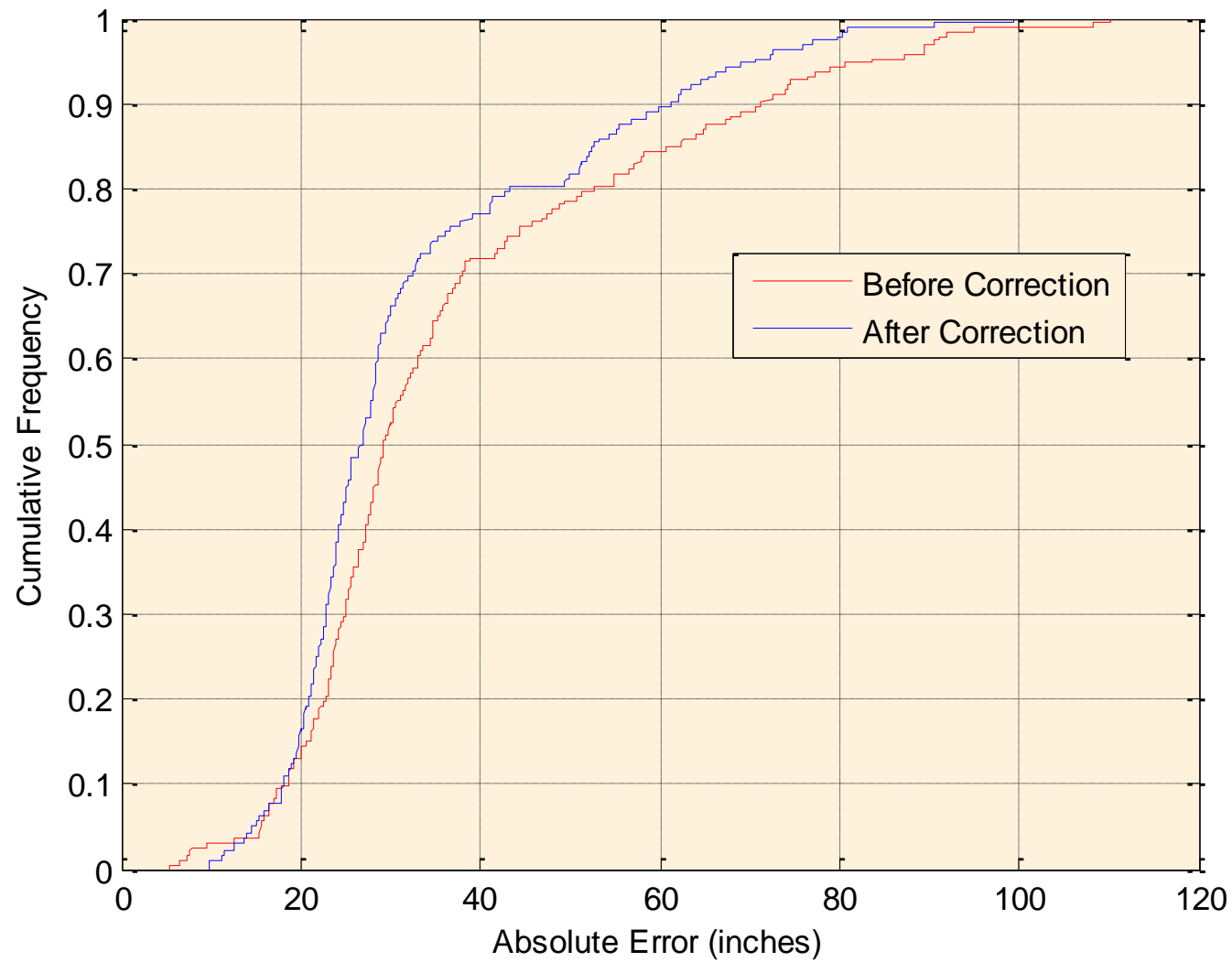
θ_j'' adjusted NEXRAD value
(2km*2km)

Bias Corrections

- Cumulative plots of precipitation depths over time based on rain gage observations and NEXRAD estimates are assessed.
- All the 192 rain gages are assessed for all the years (1996-2001). In general all the cumulative plots show good agreement between observed and NEXRAD based estimated values
- The cumulative frequencies show that there are improvements in the precipitation estimates after the bias corrections are made.

Bias Corrections





Observations

- The study reports development, implementation and evaluation of six methods for transformation of HRAP grid (4km x 4 km) based precipitation estimates to NEXRAD grid (2km x 2km) used by the SFWMD.
- Out of the six methods developed, the best method based on the evaluation of several performance measures was selected to obtain NEXRAD grid based estimates for the year 1996 till 2001.
- Data available at 192 gages located in the SFWMD were used in the assessment of the methods. The estimates values from the best method showed good agreement with rain gage data.

Observations(contd...)

- The selected method is recommended to generate precipitation estimates for NEXRAD grid (2km x 2km) to complete the precipitation data from 1996 to till date for the SFWMD. The precipitation estimates are improved via standard bias correction methods.
- The NEXRAD estimates for 2km x 2km grid obtained using the distance based interpolation methods are corrected using standard adjustment procedure discussed in the literature. However, more conceptually advanced adjustment procedures (i.e., Kriging) can be used to improve the estimates. Detailed analysis of adjusted NEXRAD estimates relevant to extreme events needs to be conducted

Spatial Analysis for Water Resources Modeling and Management

Methods for Analysis, Interpretation and Visualization of Spatial Data

Ramesh Teegavarapu, Ph.D., P.E.

Associate Professor,

Director, Hydrosystems Research Laboratory (HRL)

<http://hrl.fau.edu>

Department of Civil, Environmental and Geomatics Engineering,
Florida Atlantic University, Boca Raton, Florida, 33431, USA

Permission to use.

- The material in the presentation is obtained from several copyright protected sources (including journal publications, books and published articles, technical presentations by author and his co-authors). Permission to use the material in this presentation elsewhere needs to be obtained from the author(s) of this presentation as well as publishing agencies which own the copyright permissions for the figures and illustrations.
- Material in the presentation can only be for Academic Use only.
- Journal articles for personal use can be obtained from author: rteegava@fau.edu
- Some of figures are not yet published in any article by the author.
- Any additional information please contact the author :
rteegava@fau.edu

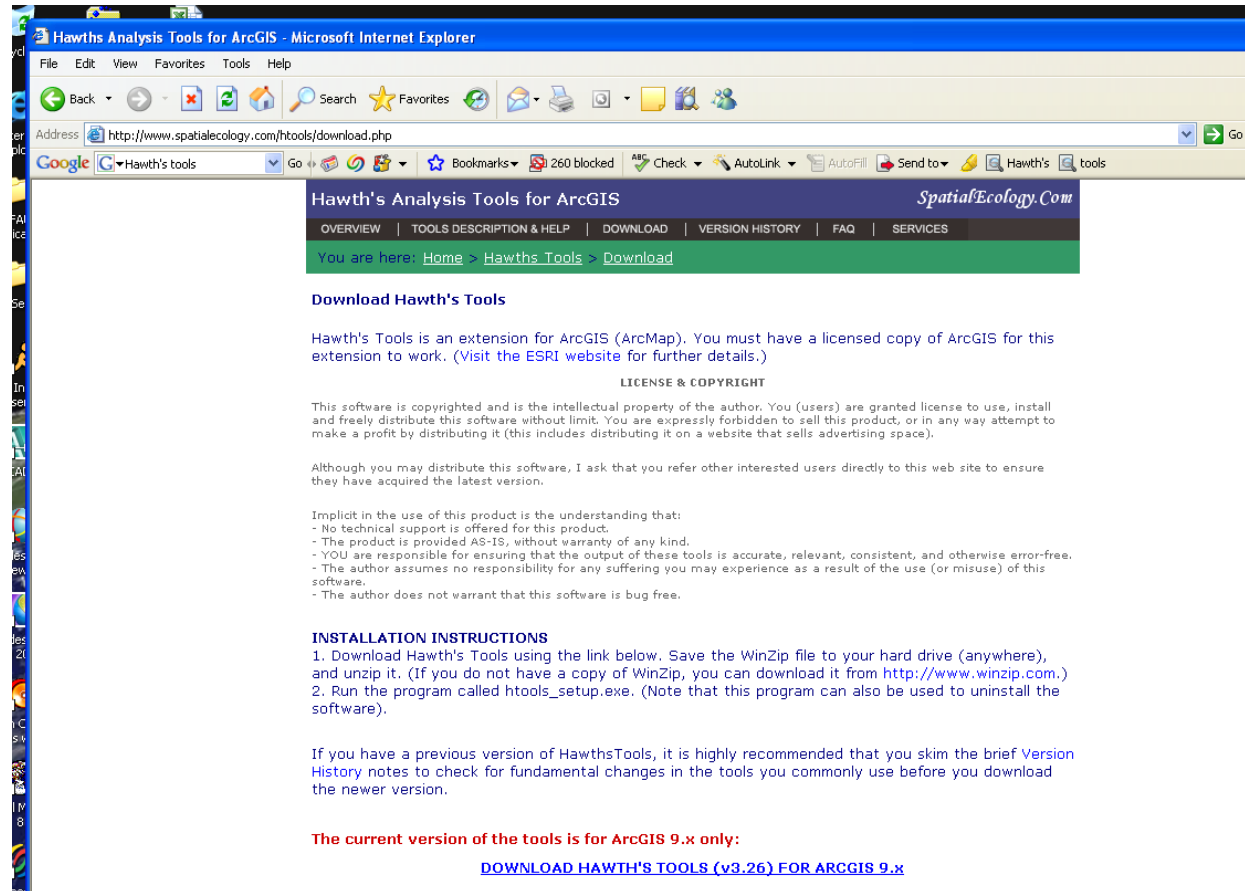
Spatial Data Analysis With GIS

- Topics
 - Introduction to geographical information.
 - Types of data, database and attributes,
 - Spatial queries
 - Data models, maps and projections, coordinate systems, representations of Earth,
 - Raster and vector data,
 - Processing of raster and vector data, spatial estimation, developing spatial data.
 - Digital elevation models, terrain analysis.
 - Concepts of digital elevation model processing
 - Geoprocessing of hydrometeorology datasets.

GIS Software

- ArcGIS [different versions]
- Hawth's Tools (for Analysis and Calculations)
– Add-in to ArcGIS – Public Domain.
- ArcGIS software with limited time (one year)-full version is generally available to university students if the university has an agreement with ESRI for university-wide license.
- The software given to students can be used for ACADEMIC purposes only.

Hawth's Tools for ArcGIS

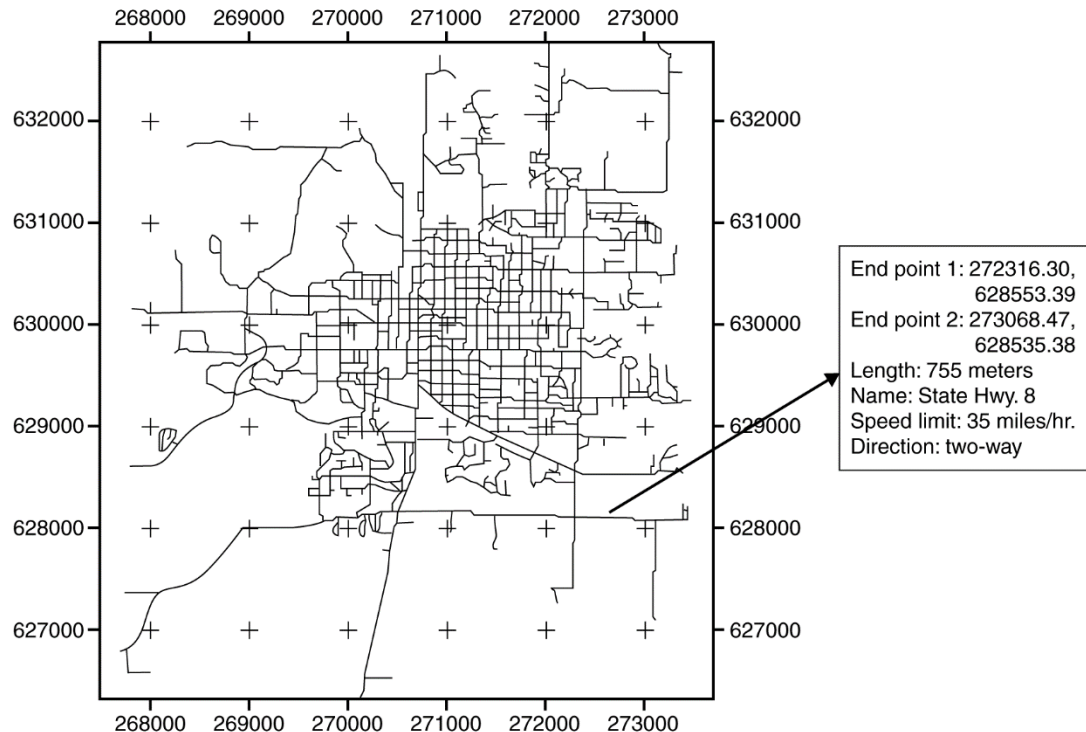


- <http://www.spatial ecology.com/htools/download.php>

GIS ?

- A geographic information system (GIS) is a computer system for **capturing, storing, querying, analyzing, and displaying** geospatial data.
- Geospatial technology is listed by the U.S. Department of Labor as one of the three emerging industries, along with nanotechnology and biotechnology.
- <http://www.careervoyages.gov/>

Example of Geographically Referenced Data



An example of geographically referenced data.
The street network is based on a plane coordinate system.
The box on the right lists the x- and y-coordinates of the end points and other attributes of a street segment.

Spatial Analysis Tools

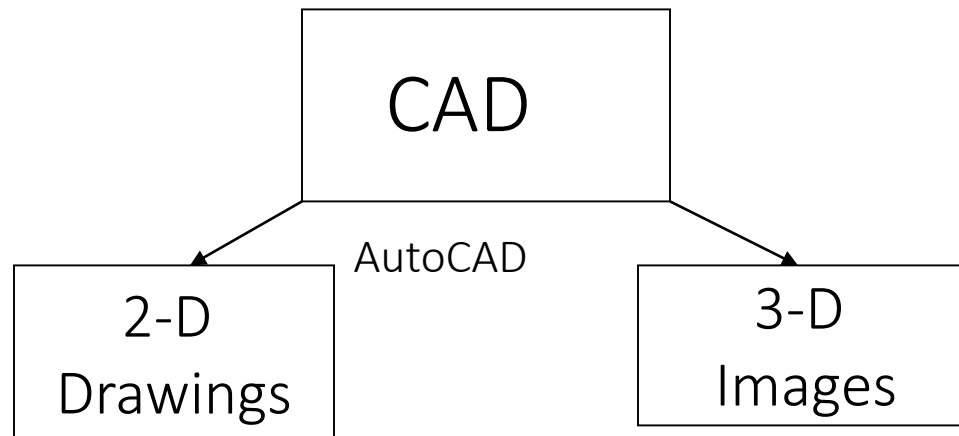
- Graphical Tools
- GIS Applications
- Data Types
- Data Formats
- Spatial Coordinate Systems/Issues

Graphical Tools for Engineering

- Computer Aided Design (CAD)
- Automated Mapping/Facilities Management (AM/FM)
- Geographical Information System (GIS) – Spatial Analysis Tools

Computer Aided Design

- A system of hardware and software for use in developing, displaying, and outputting engineering drawings and graphical images.



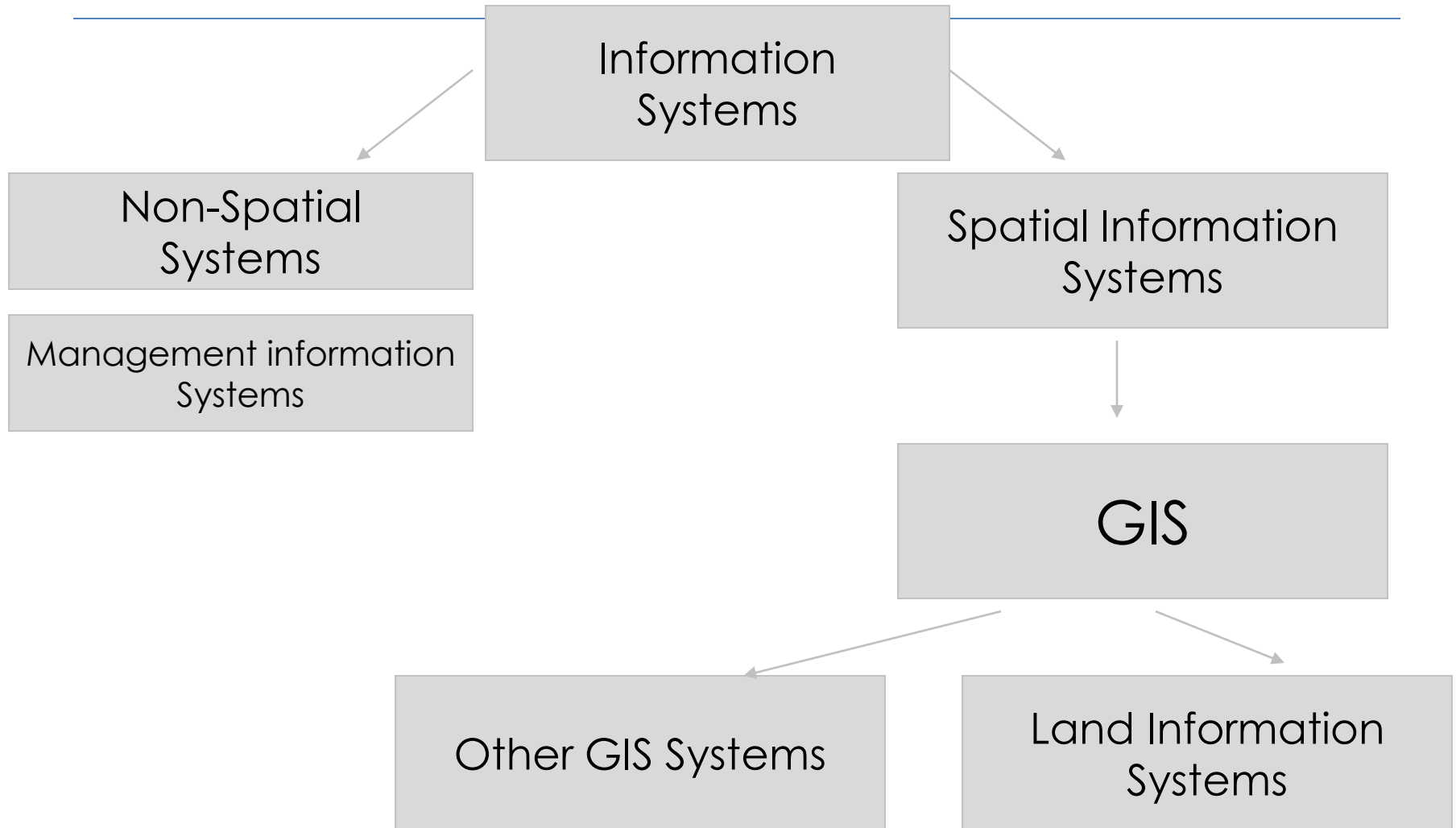
Automated Mapping/Facilities Management

- An organized collection of computer hardware, software, and data for use in visualizing and managing various components of engineering systems (e.g. traffic systems, building systems, utility systems).



Microstation

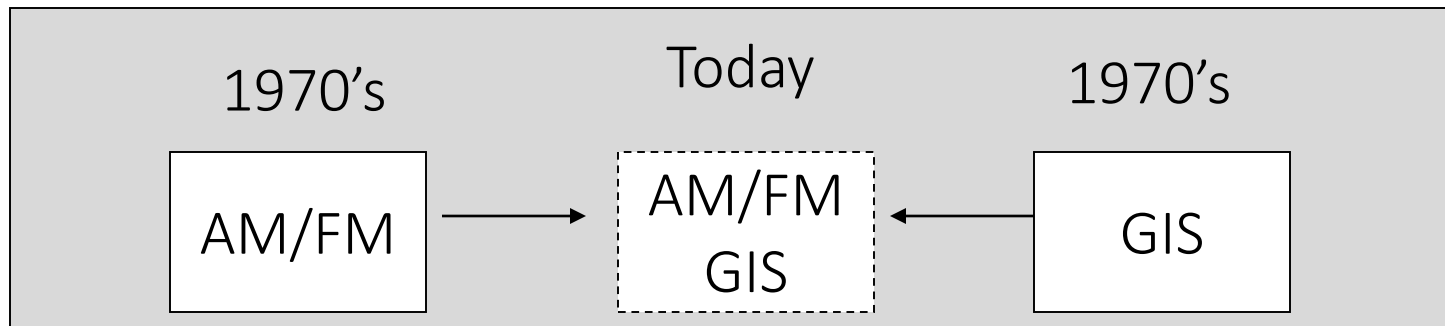
A **taxonomy** of Information Systems



GIS Subsystem Functions

- A data input subsystem that collects and preprocesses spatial data from various sources.
- A data storage and retrieval subsystem that organizes the spatial data in a manner that allows retrieval, updating and editing.
- A data manipulation and analysis subsystem that performs tasks on the data, aggregates and disaggregates, estimates parameters and constraints and performs modeling functions.
- A reporting subsystem that displays all or part of the database in tabular, graphic or map form

AM/FM/GIS Summary

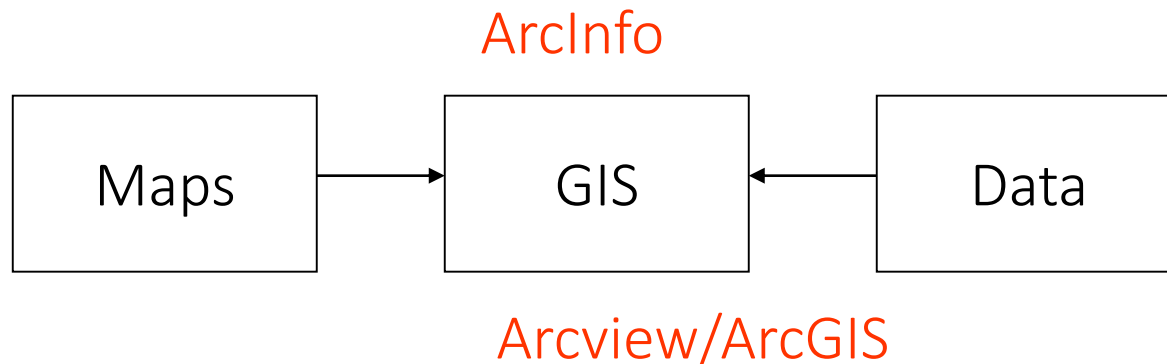


Engineers
Small Scale
Data/Drawings
Design/Management

Planners
Large Scale
Data/Maps
Planning/Analysis

Geographic Information System

- An organized collection of computer hardware, software, and data designed to capture, store, update, manipulate, analyze, display, and output all forms of geographically referenced information.



AM/FM/GIS Applications

- Presentation and Thematic Mapping
 - Show map of features
 - Show map of attributes
- Database Access
 - Input/Edit attribute data associated with features
- Database Integration
 - Link features with other attribute data

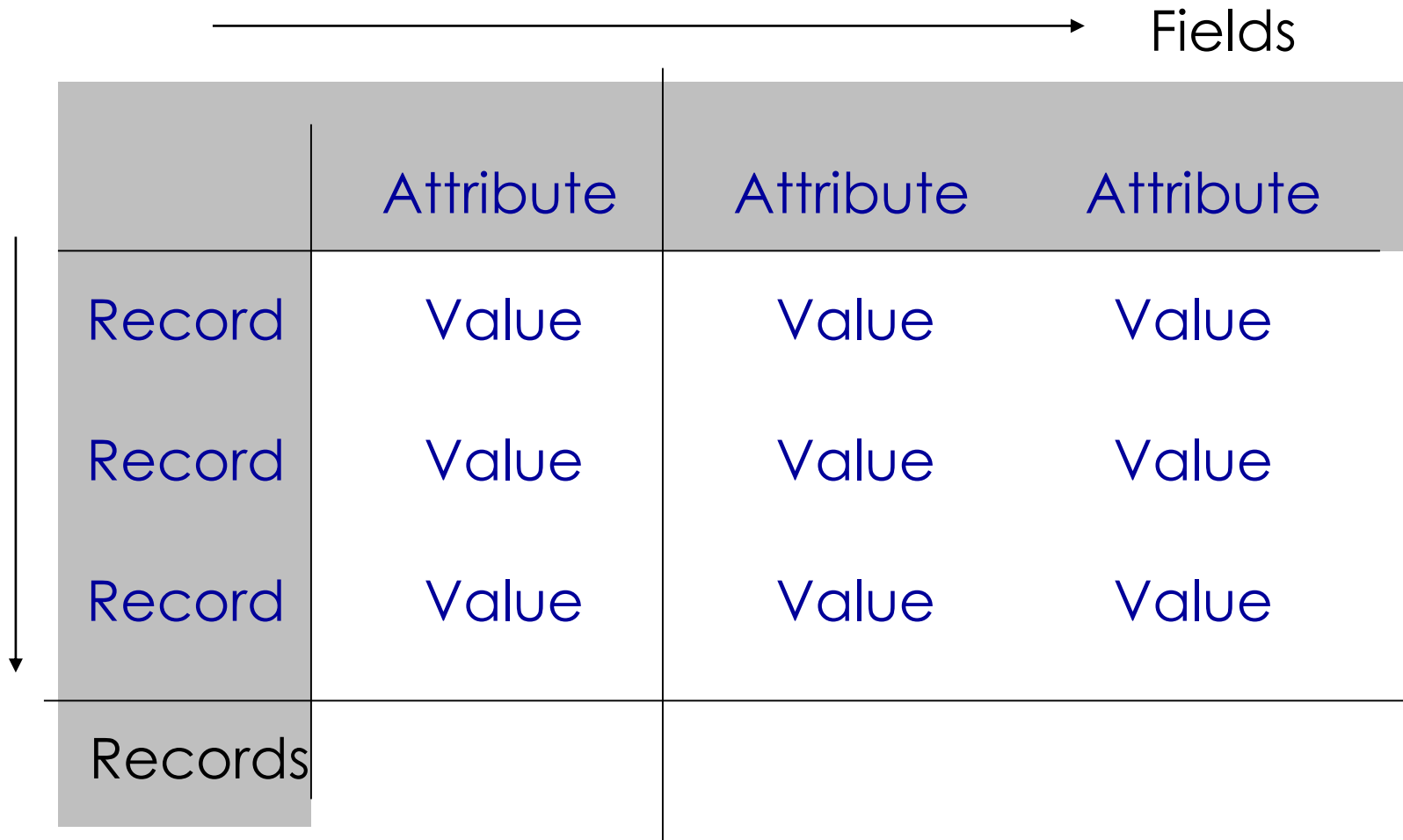
Elements of GIS

- Database (Relational)
- Geographic Information
- Computational Power to handle analysis
- Software for visualization
- Linking of information from database and spatial objects

Database

- In a database, we store attributes as column headers and records as rows.
- The contents of an attribute for one record is a value.
- A value can be numerical or text.
- Data in a GIS must contain a geographic reference to a map, such as latitude and longitude.
- The GIS cross-references the attribute data with the map data, allowing searches based on either or both.
- The cross-reference is a link.

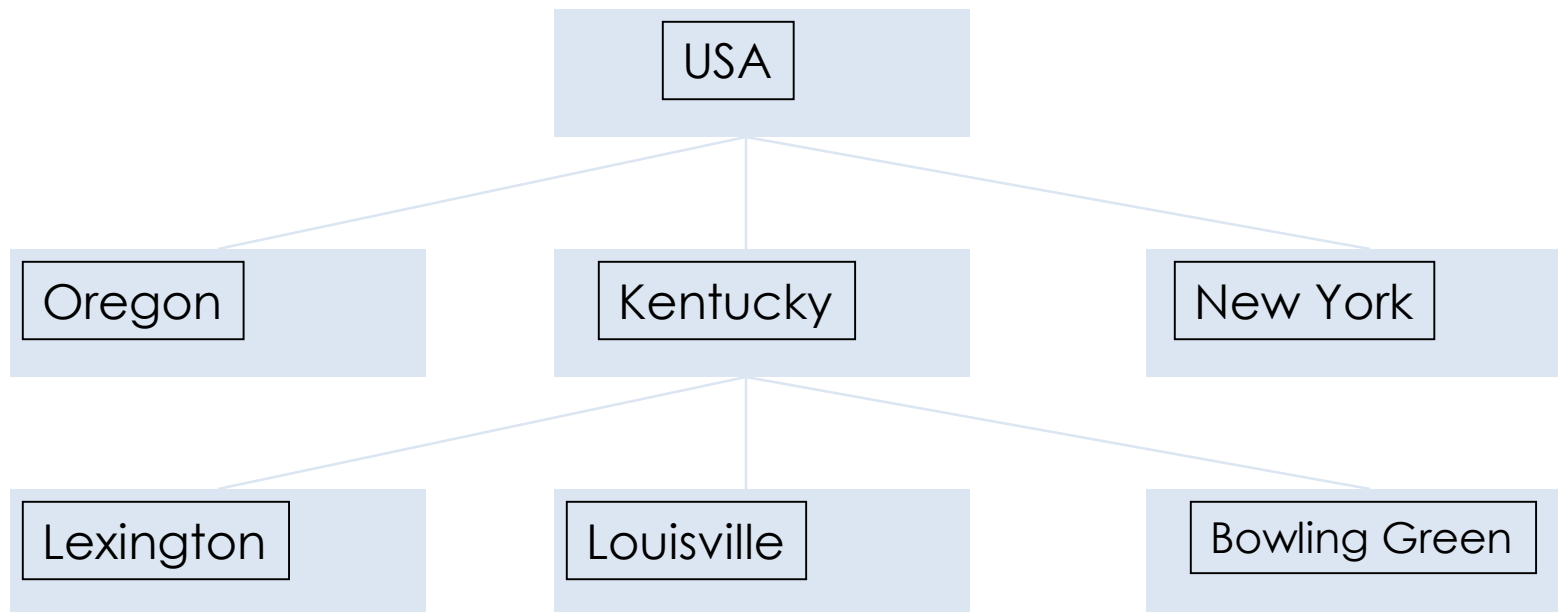
Flat File Database



The diagram illustrates a Flat File Database structure. It features a grid with a header row and three data rows. The header row is shaded gray and contains the labels 'Attribute', 'Attribute', and 'Attribute'. The data rows are white and contain the label 'Value' in each of the three columns. To the left of the grid, a vertical arrow points downwards, indicating the sequence of records. Above the grid, a horizontal arrow points to the right, indicating the sequence of fields. The word 'Fields' is positioned to the right of the horizontal arrow. The word 'Records' is positioned at the bottom left of the grid, below the first column.

	Attribute	Attribute	Attribute
Record	Value	Value	Value
Record	Value	Value	Value
Record	Value	Value	Value
Records			

Hierarchical Database



Relational Database

Patient Record

File

Key	Check-in	Check Out	Room No.
42	2/1/96	2/4/96	N763
78	2/2/96	2/4/96	N712

Purchase Record

File

Item	Date	Price	Customer	Key
Skate Board	2/1/96	49.95	John Smith	42
Baseball Bat	2/1/96	17.99	James Brown	778

Accident Report

File

Date	Injury	Name	Key	Location
2/1/96	Broken Leg	John Smith	42	75 Elm Street
2/2/96	Concussion	Sylvia Jones	654	12 State Street
2/2/96	Cut on Ear	Robert Doe	123	2323 Broad Street

Example by Clarke

AM/FM/GIS Applications

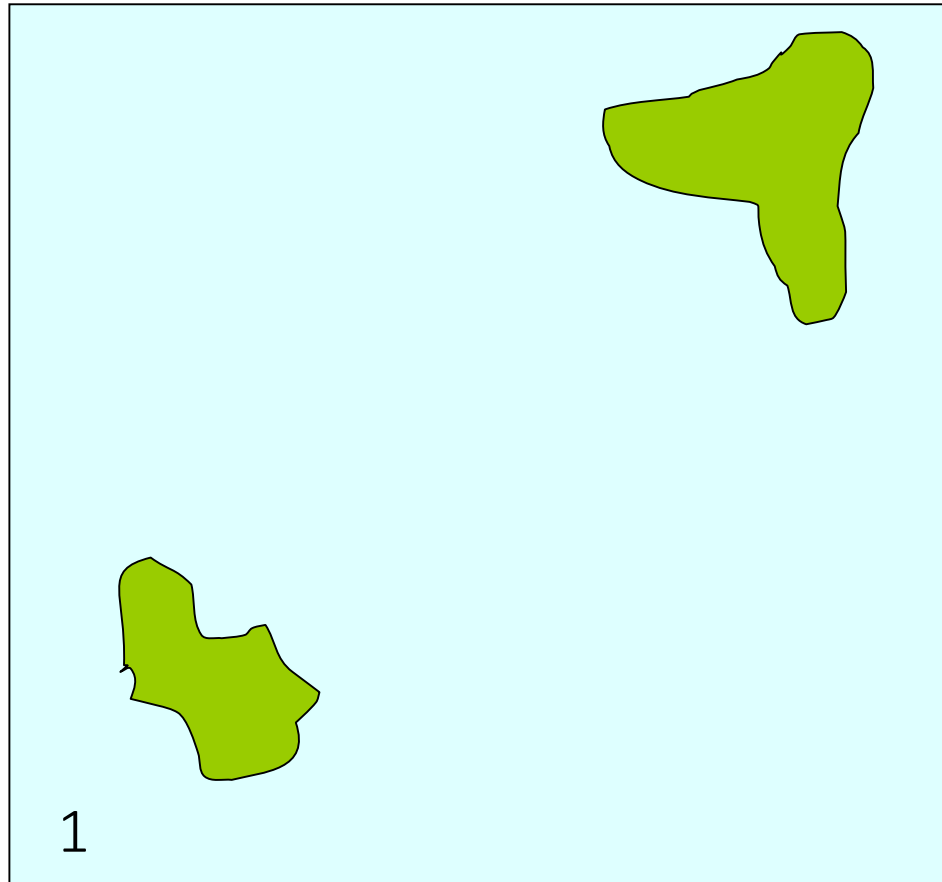
- Data Query
 - Show location of features with certain attributes
- Distance Analysis
 - Calculate map distance between objects
 - Calculate network distance between objects
- Routing and Minimum Path Analysis

AM/FM/GIS Applications

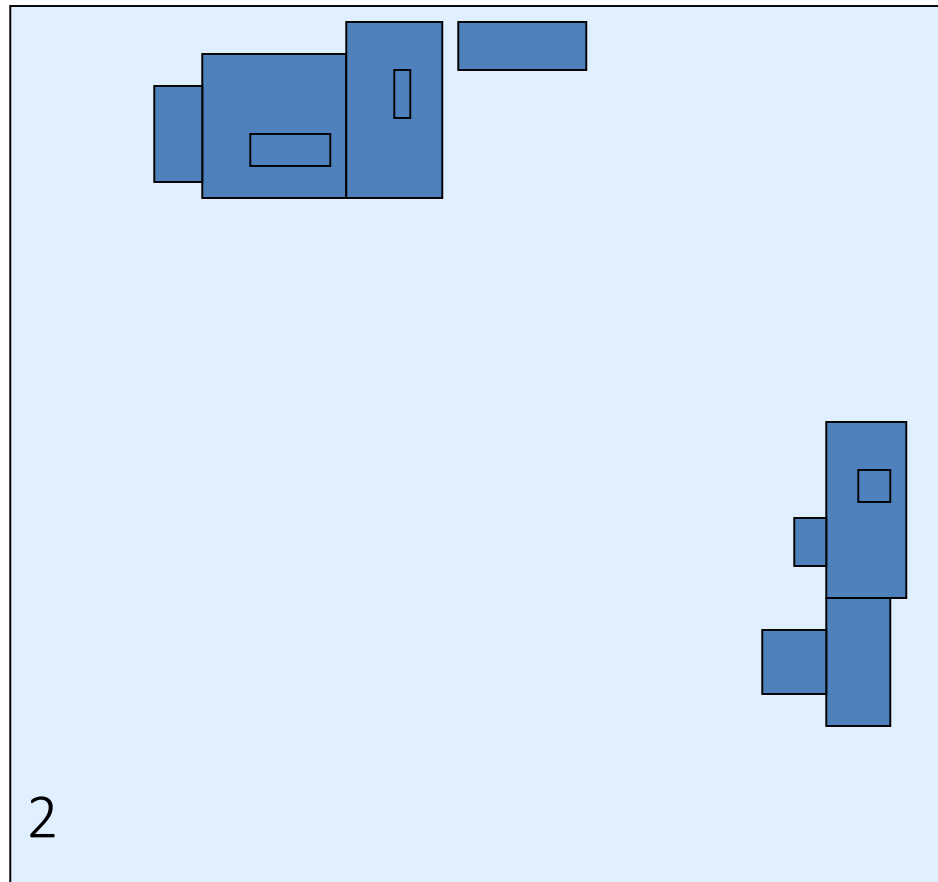
- Presentation and Thematic Mapping
- Database Access
- Database Integration
- Database Query
- Distance Analysis
- Routing and Minimum Path Analysis

A simple Illustration of how GIS works

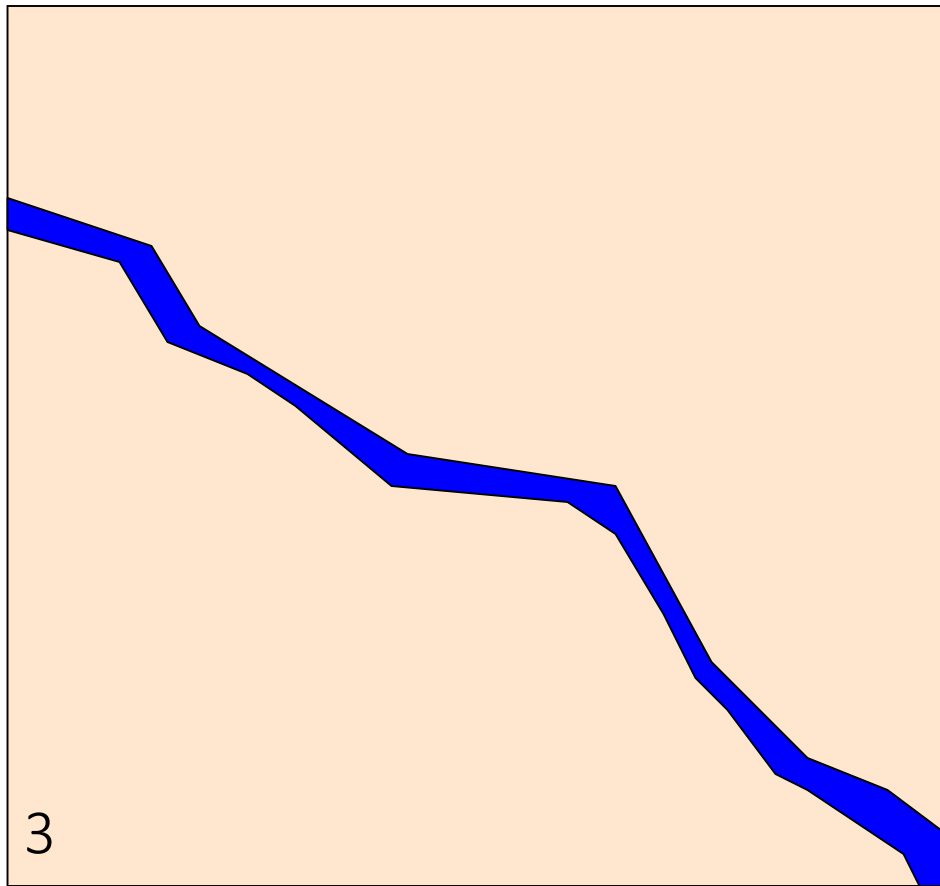
Map of parks



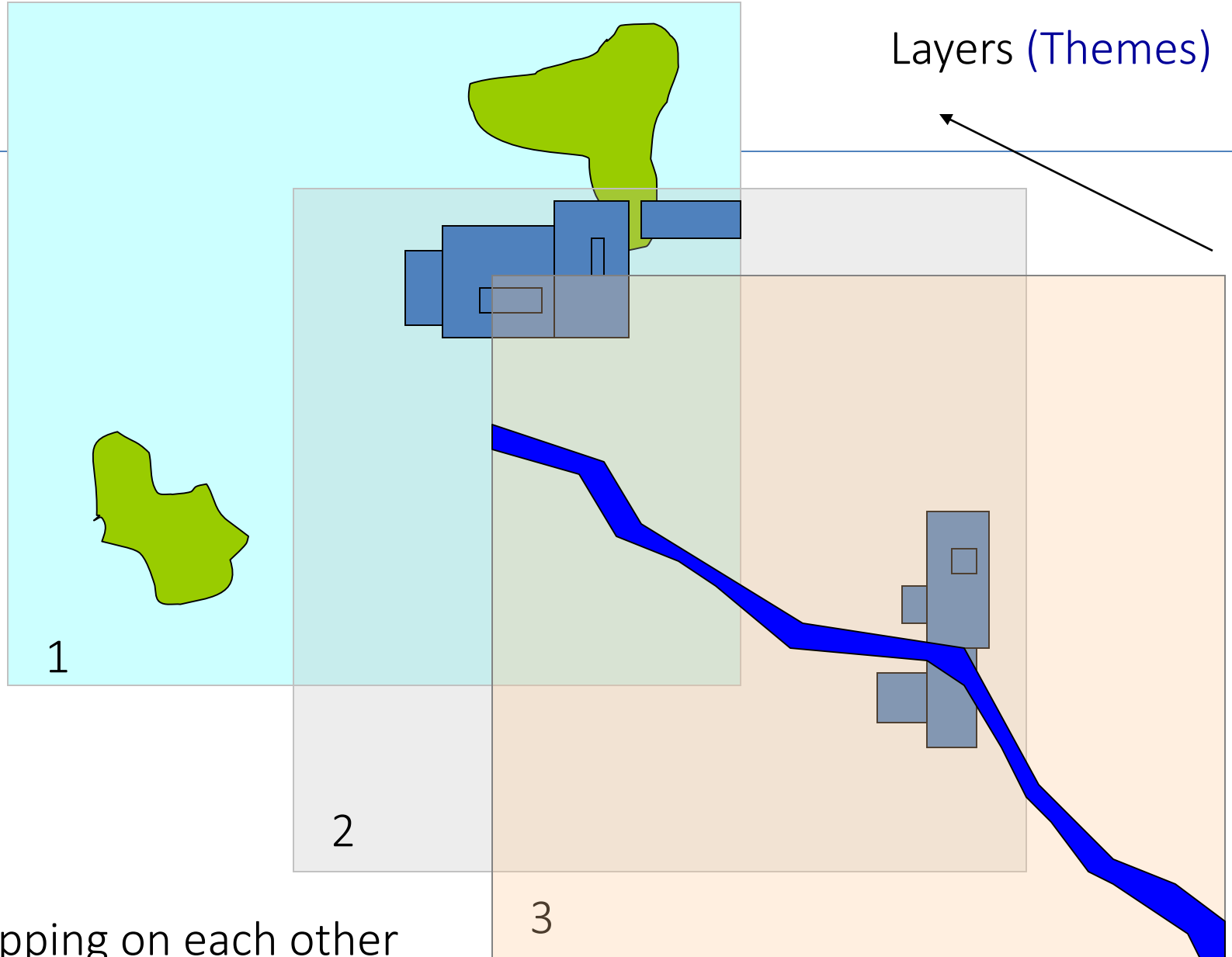
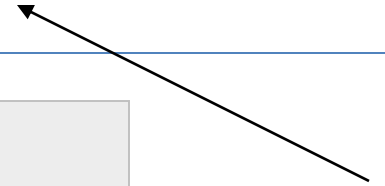
Map of Residential Areas



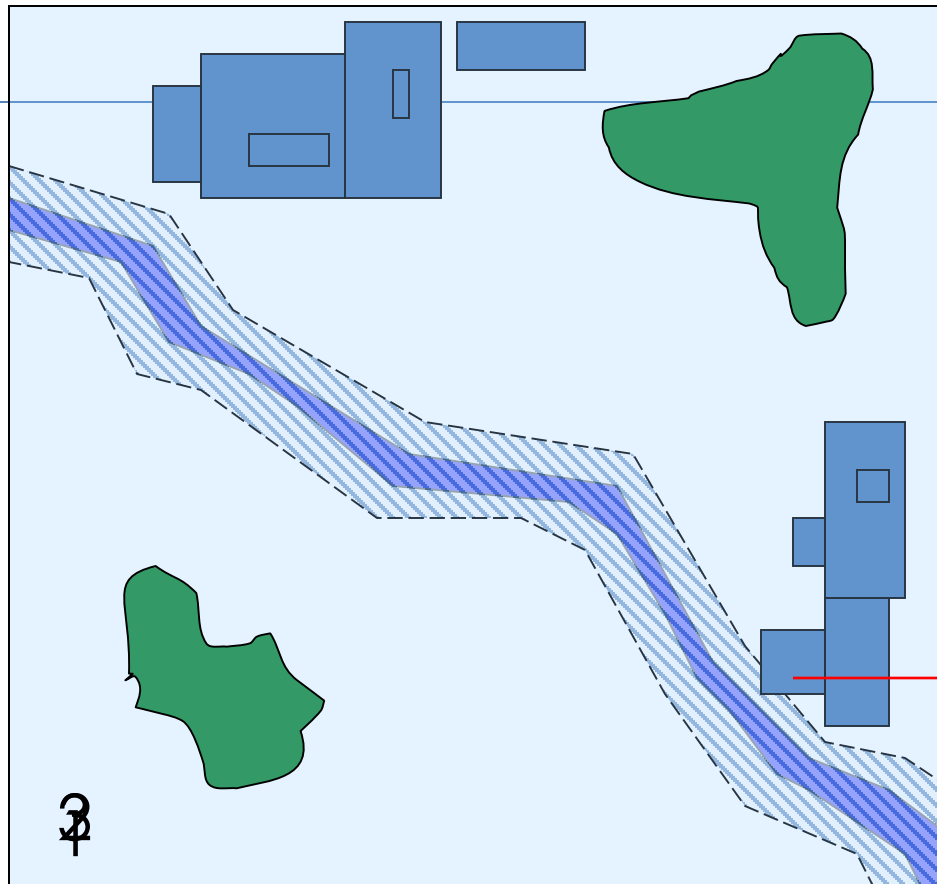
Map of River



Layers (Themes)



Over lapping on each other



Creation of
Flood Buffer
Zone around
river

This
Building is
in the new
Buffer zone

All Maps Over layed on top of each other. Buffer zone
Around the river is created using GIS software

GIS Applications

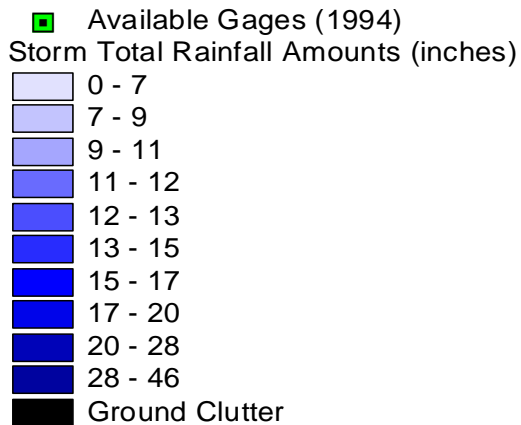
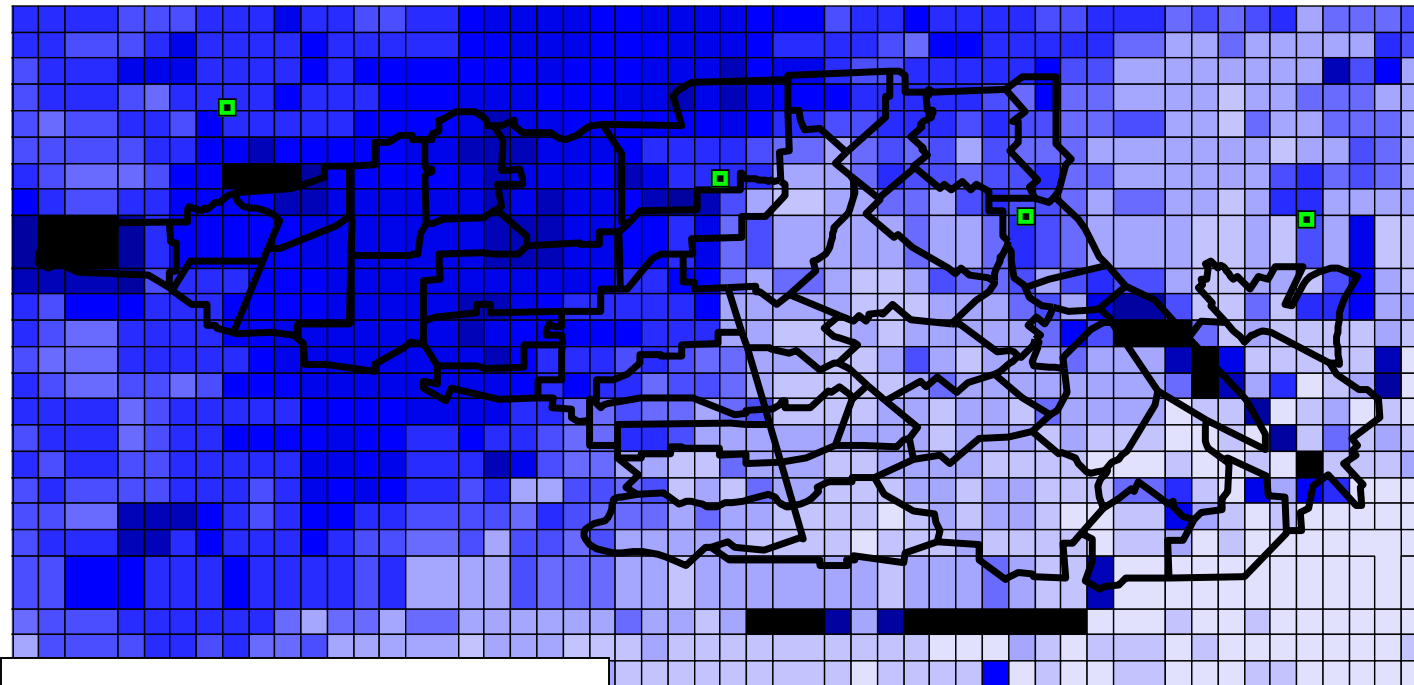
- Spatial Query
 - Show location of features with certain attributes within a certain area
- Adjacency Analysis
 - Determine which map features touch or are adjacent to other map features.
- Proximity Analysis
 - Determine features that are near to other map features
- Buffer Analysis
 - Creating buffer around features.

GIS in Hydrology

- **Spatial character** of various hydrologic parameters in the hydrologic processes
- Spatial variability of **rainfall**, **infiltration**, **land use/land cover**, **hydraulic roughness**, **slope**, ... etc.
- Digital representation of these processes required spatial analysis techniques.
- GIS is valuable in watershed delineation, runoff estimation, hydraulic modeling and flood plain mapping.

NEXRAD Data

October 14th - 19th 1994 Storm Totals

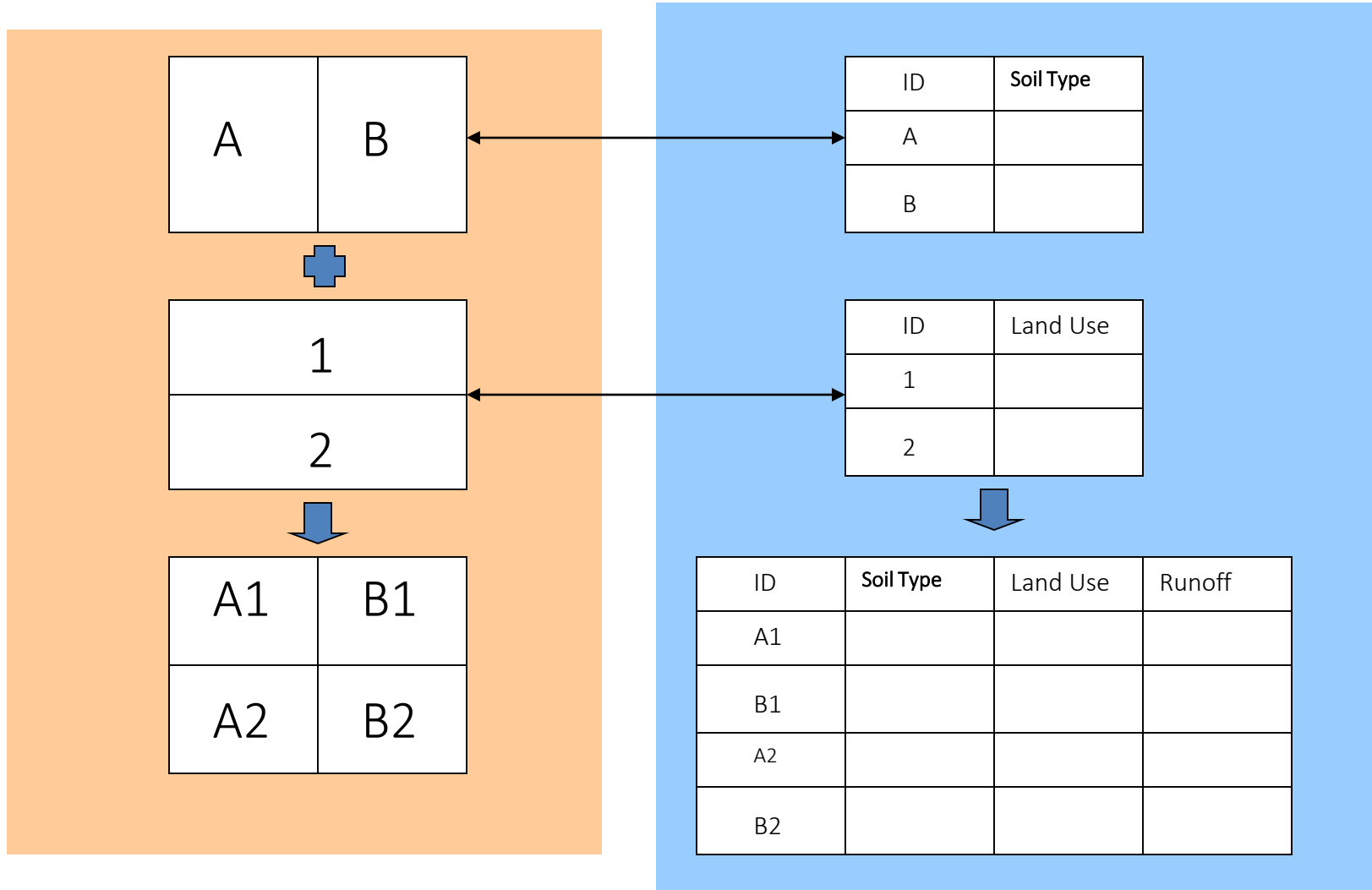


Example use of spatially distributed rainfall
in water resources modeling by using a grid
of values

Example of a Query Application

- Determine possible sites for a reservoir:
 - Watershed Area > 10 sqmi
 - No fault lines near dam site
 - No major roads inundated
 - No homes or major utilities submerged
 - No homes within 1 mile downstream

Example Overlay Analysis (hydrology)



GIS – Time tested Definition

"A geographic information system is a special case of information systems where the database consists of observations on spatially distributed features, activities or events, which are definable in space as points, lines, or areas. A geographic information system manipulates data about these points, lines, and areas to retrieve data for ad hoc queries and analyses" (Dueker, 1979).

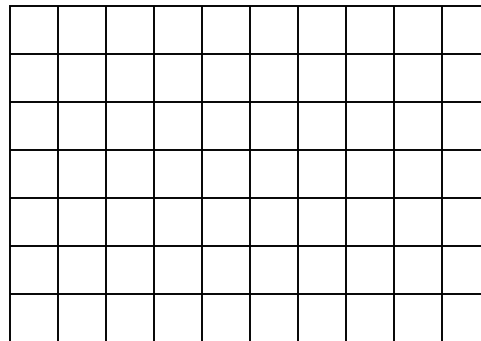
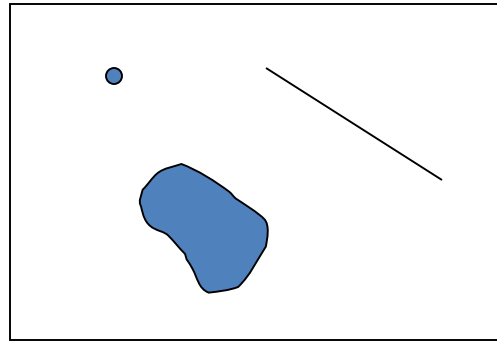
Representations of real-world

GIS Data Types

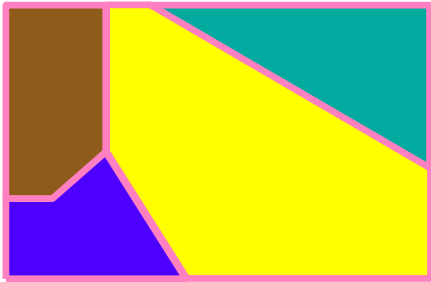
- Spatial Data
 - Location of geographic features
- Image Data
 - Graphical representation of objects
- Tabular Data
 - Descriptive information usually related to geographic features

Spatial Data Types

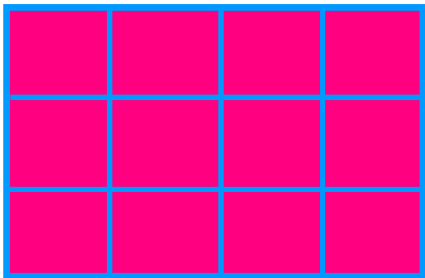
- Vector Data
 - Points
 - Lines
 - Polygons
- Raster Data



Discrete and Continuous Space



Discrete Space: Vector GIS
Lumped models

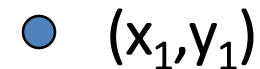


Continuous Space: Raster GIS, TIN
Distributed models

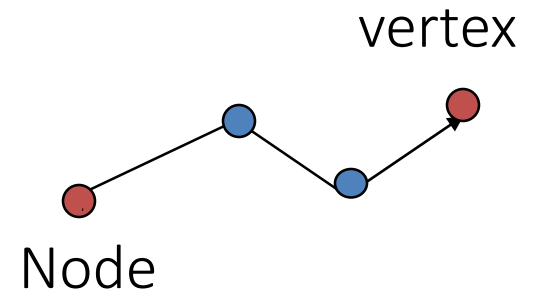
Spatial Data: Vector format

Vector data are defined spatially:

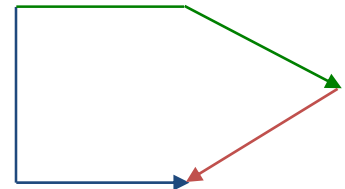
Point - a pair of x and y coordinates



Line - a sequence of points

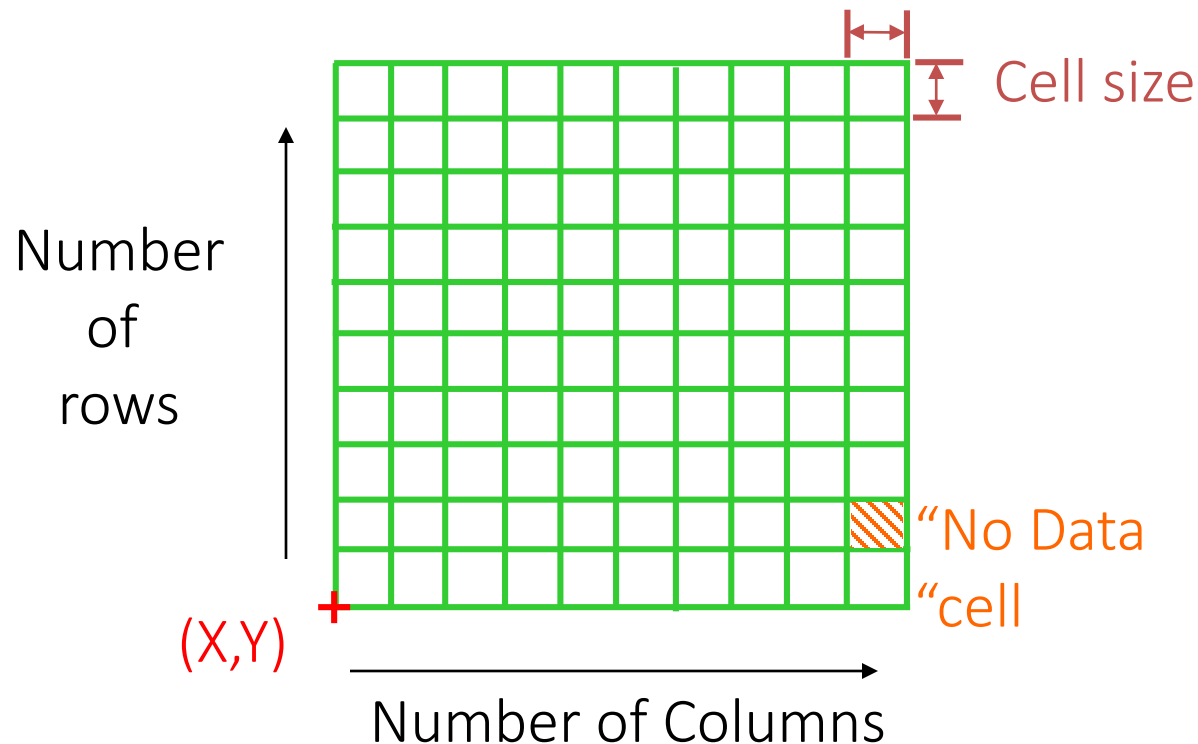


Polygon - a closed set of lines

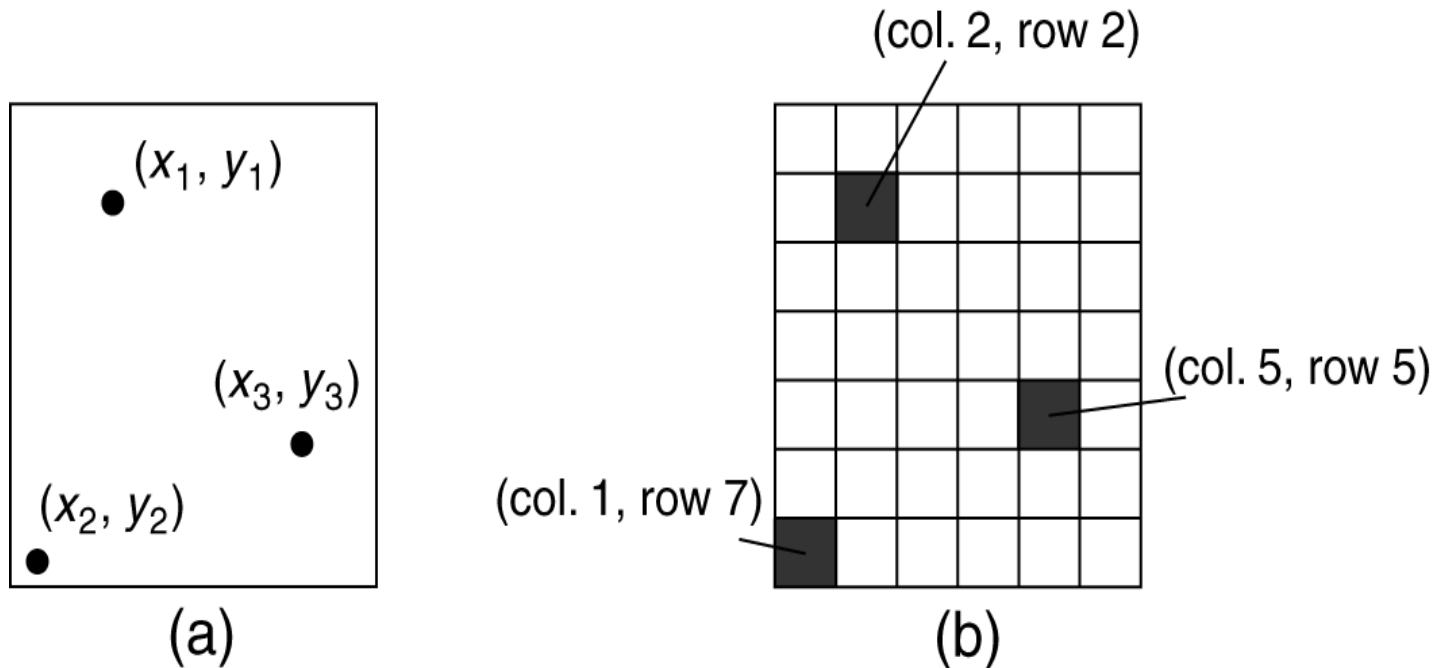


Spatial Data: Grid (Raster) format

Raster data are described by a cell grid, one value per cell:

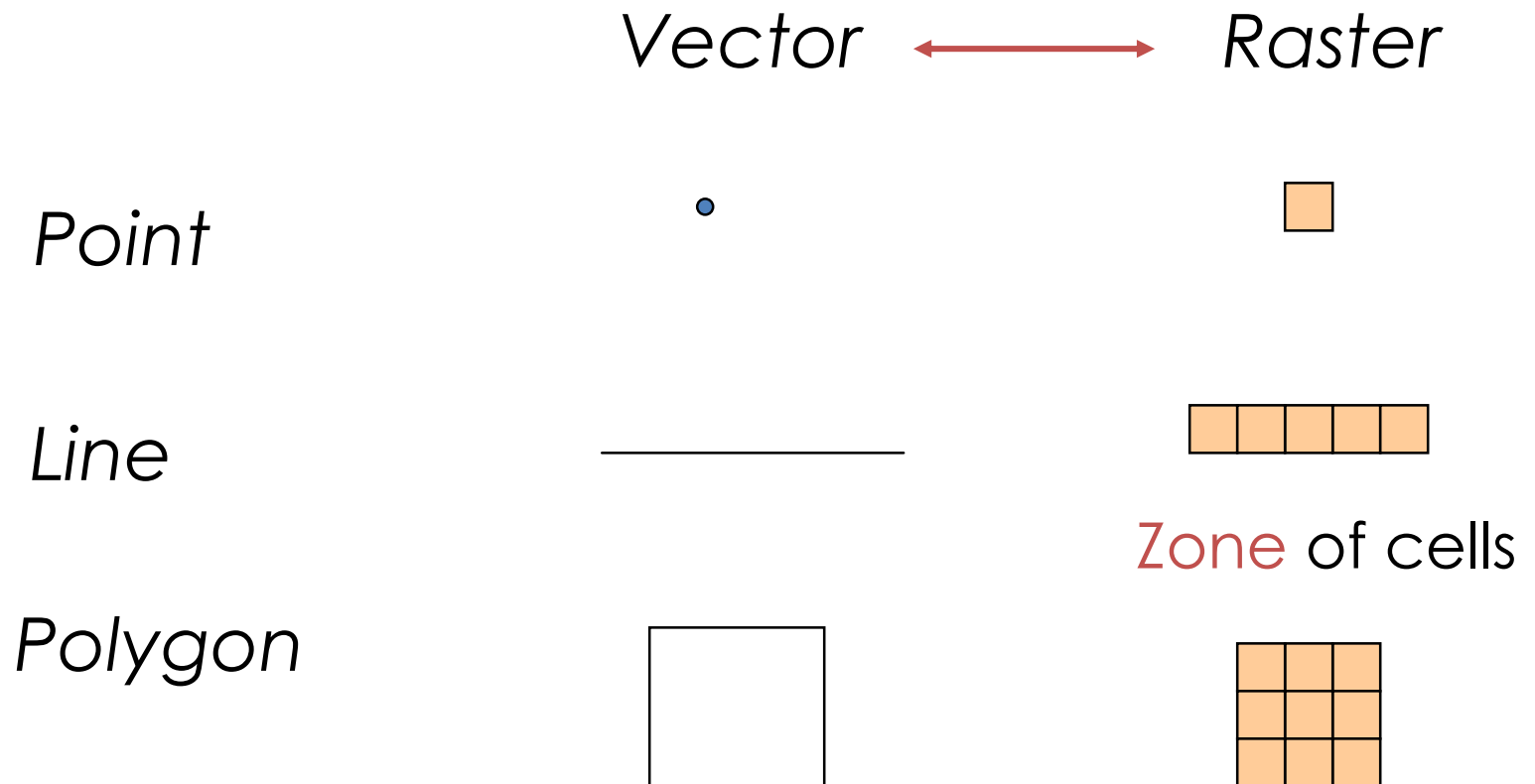


Vector and Raster Representation



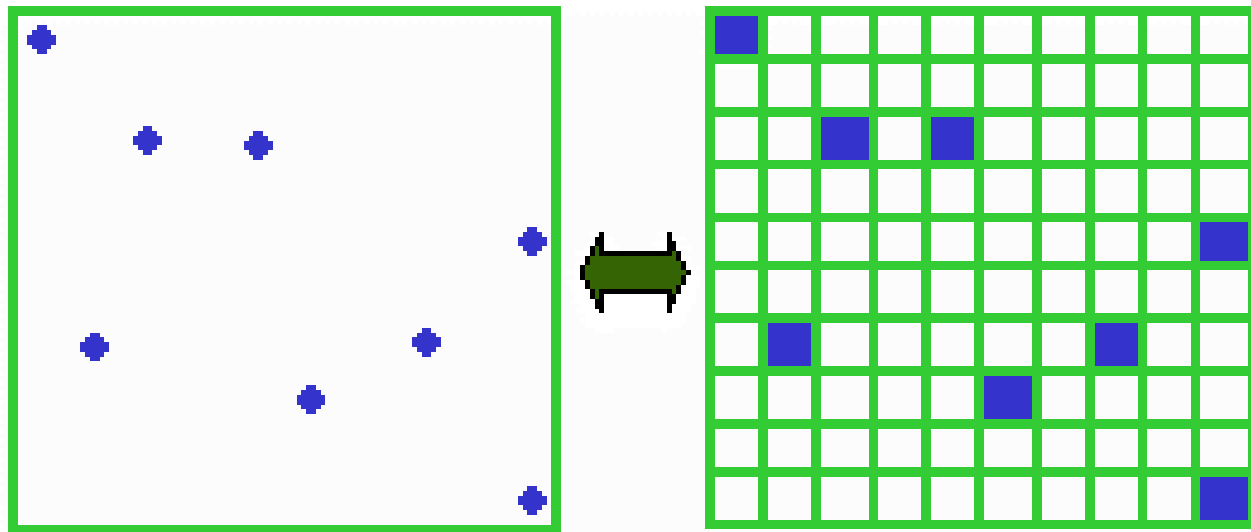
The vector data model uses x-, y-coordinates to represent point features (a), and the raster data model uses cells in a grid to represent point features (b).

Raster and Vector Data



Points represented as Cells

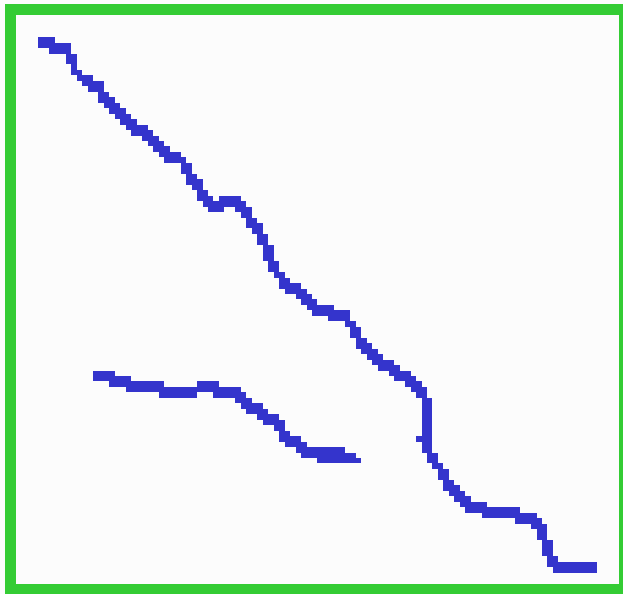
No Dimensions



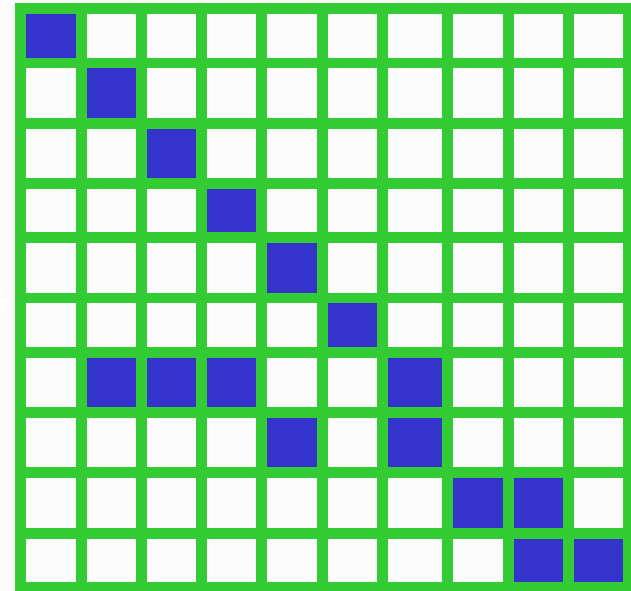
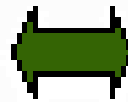
Vector to raster

Line represented as a Sequence of Cells

one Dimension



Vector

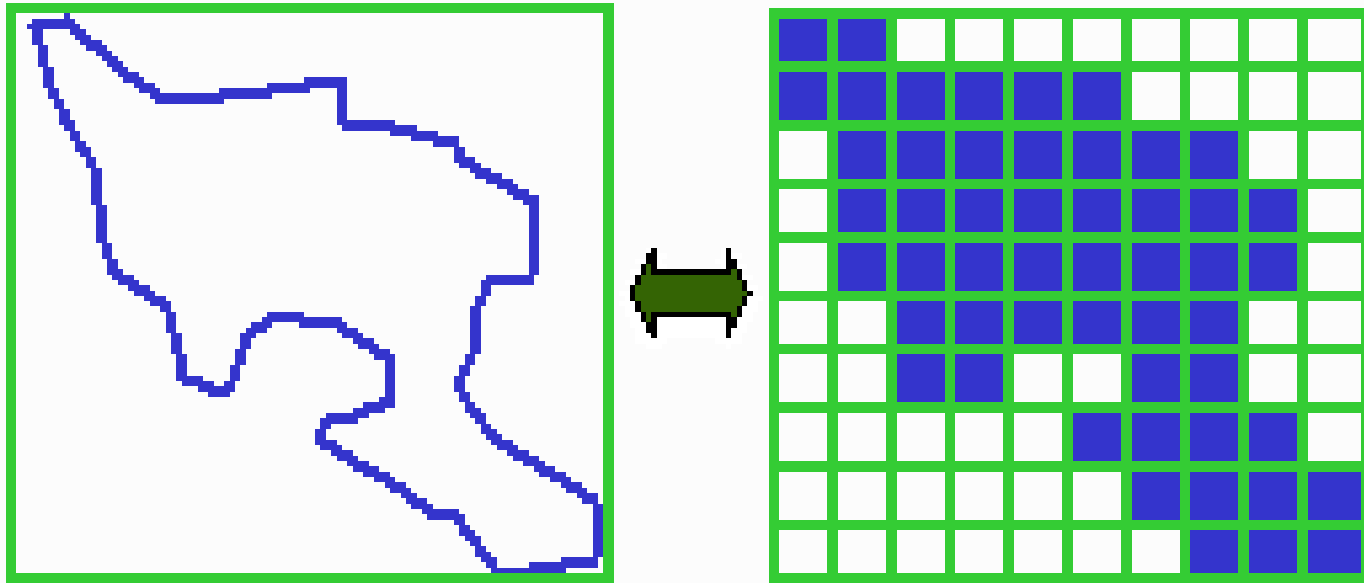


Raster

Vector to raster

Polygon as a Zone of Cells

Two Dimensions



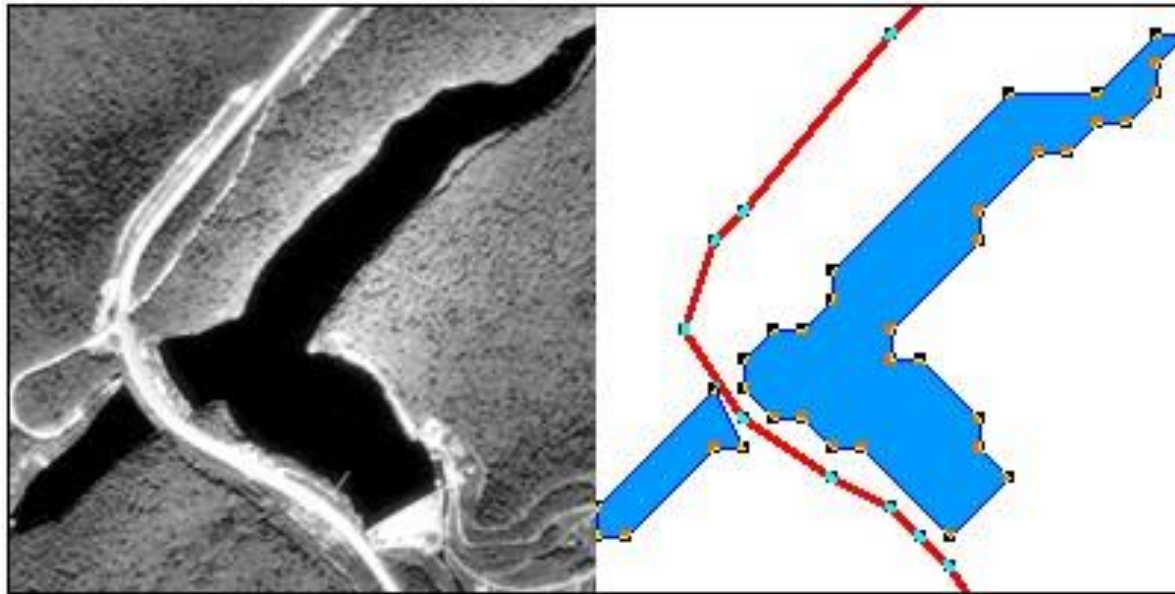
Vector

Raster

Vector to raster

Vector Data

- Uses positions to represent real world entities
 - Points, lines, polygons

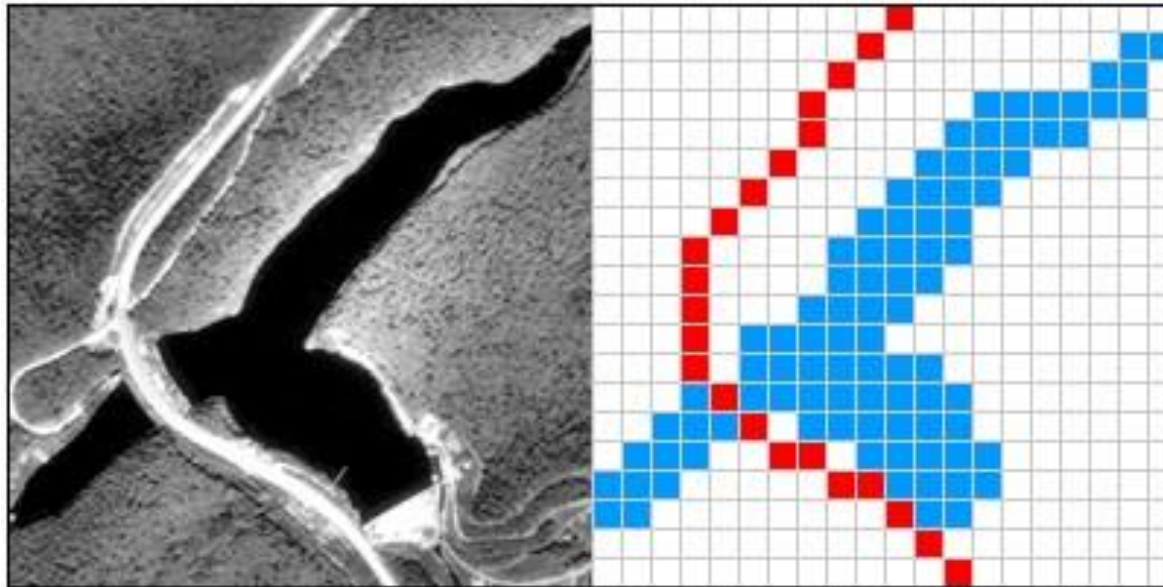


Reservoir and Highway

Raster to vector

Raster Data

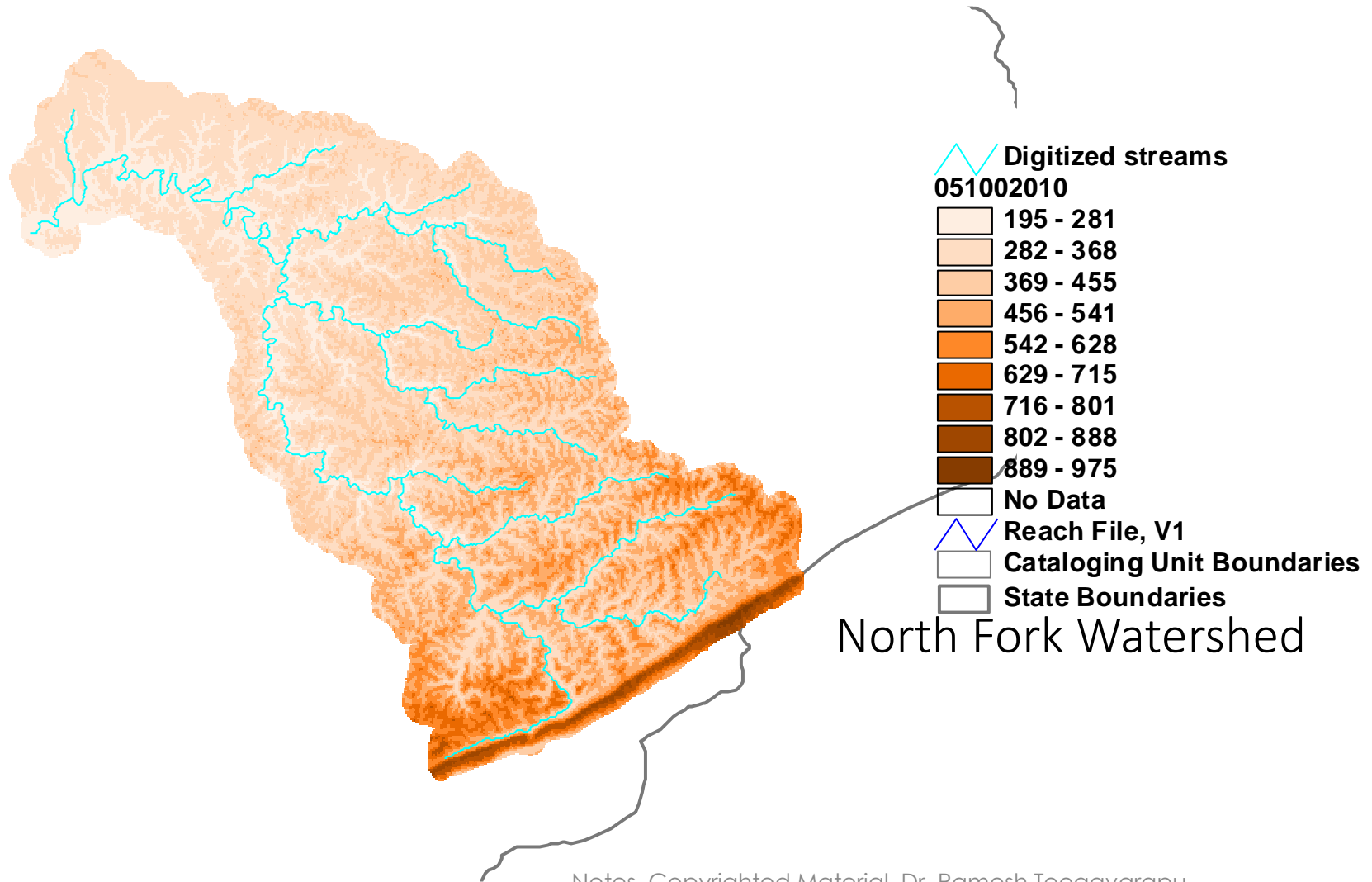
- Samples attributes at fixed intervals
 - one value per cell



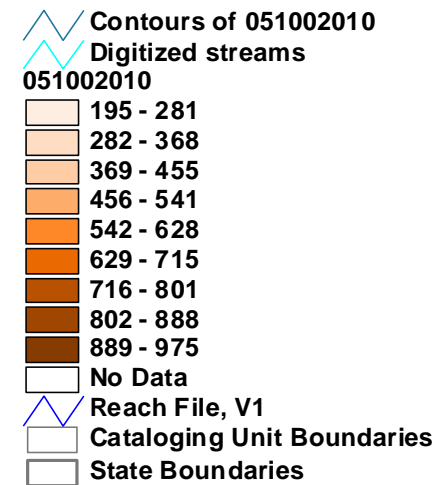
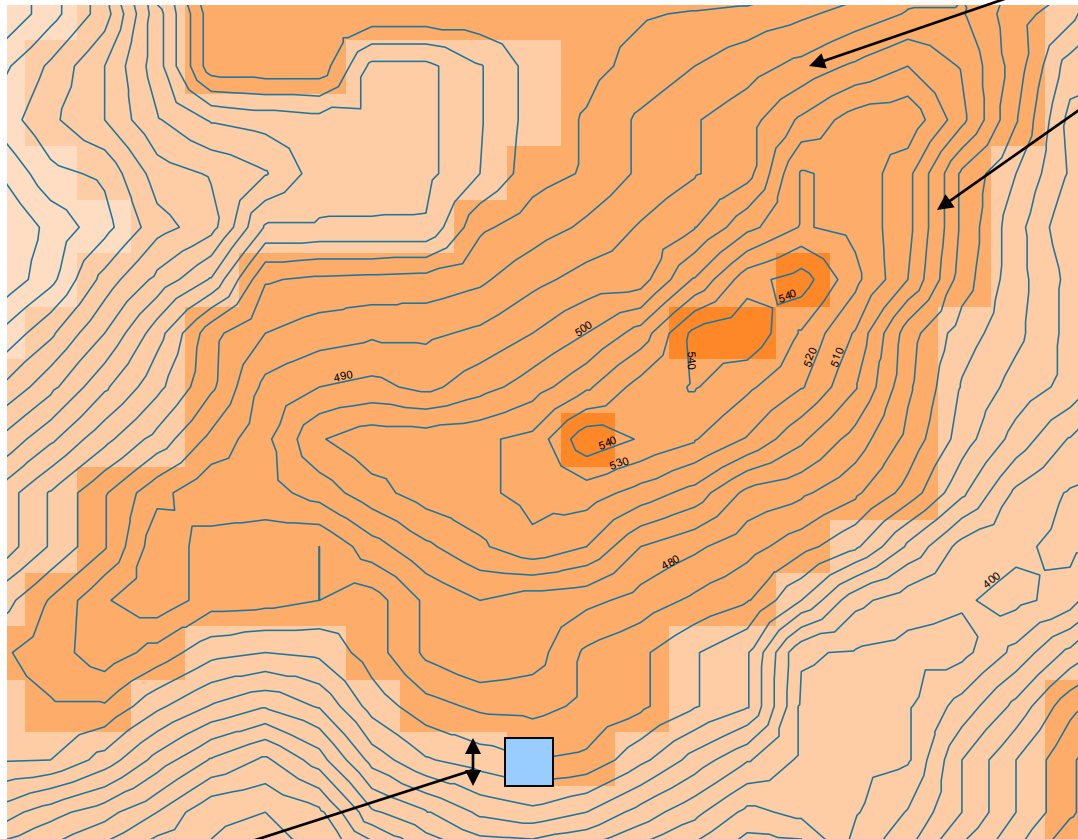
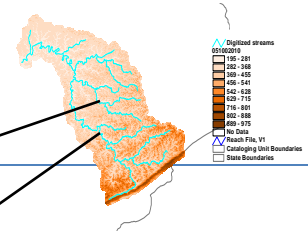
Reservoir and Highway

Digital Elevation Model (DEM)

Raster DATA



DEM - resolution

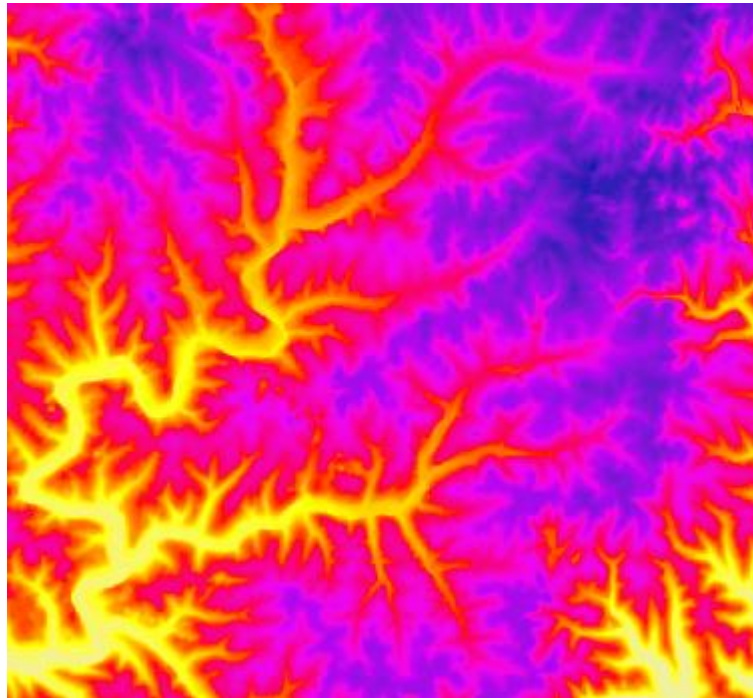


30 m X 30 m

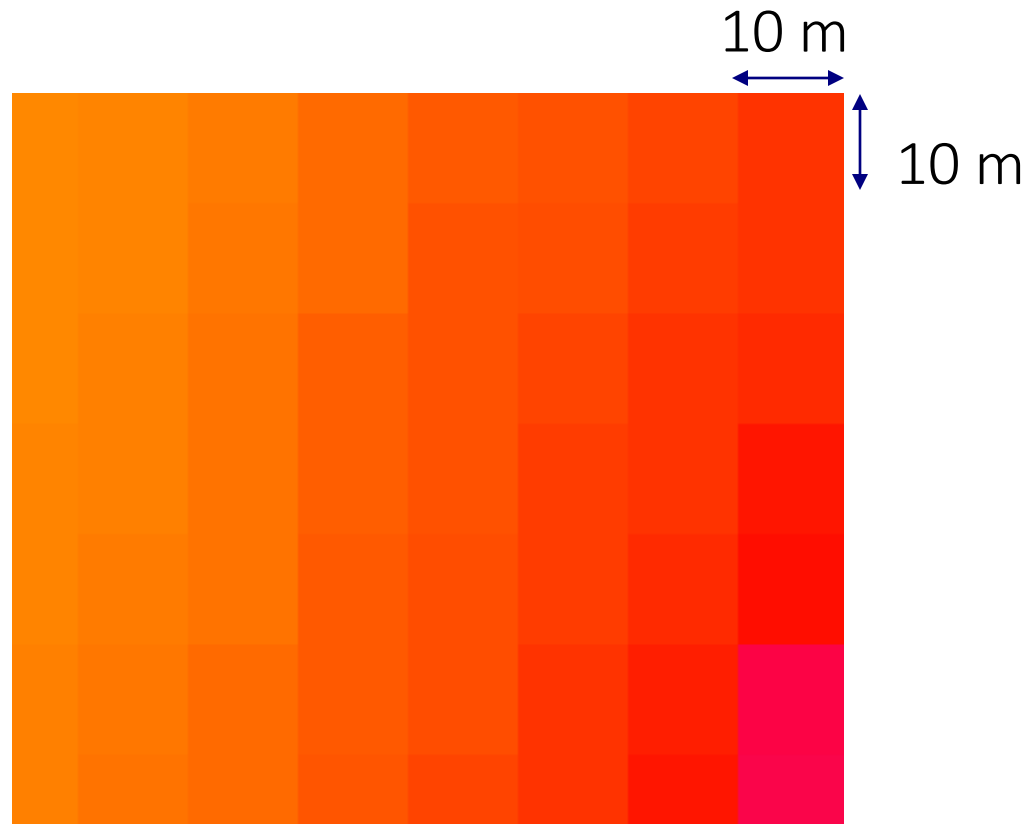
DEM – RASTER with Contours drawn

DEM

- Digital Elevation Models

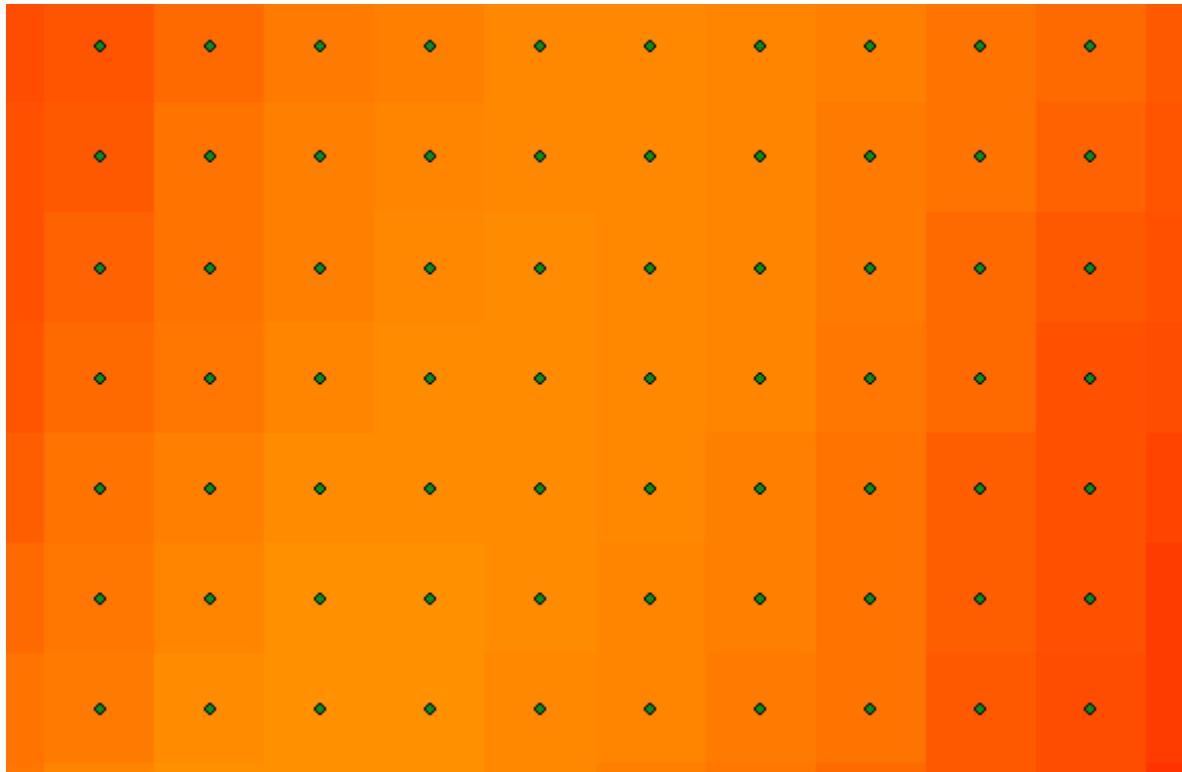


Grid Size = 10 m x 10 m

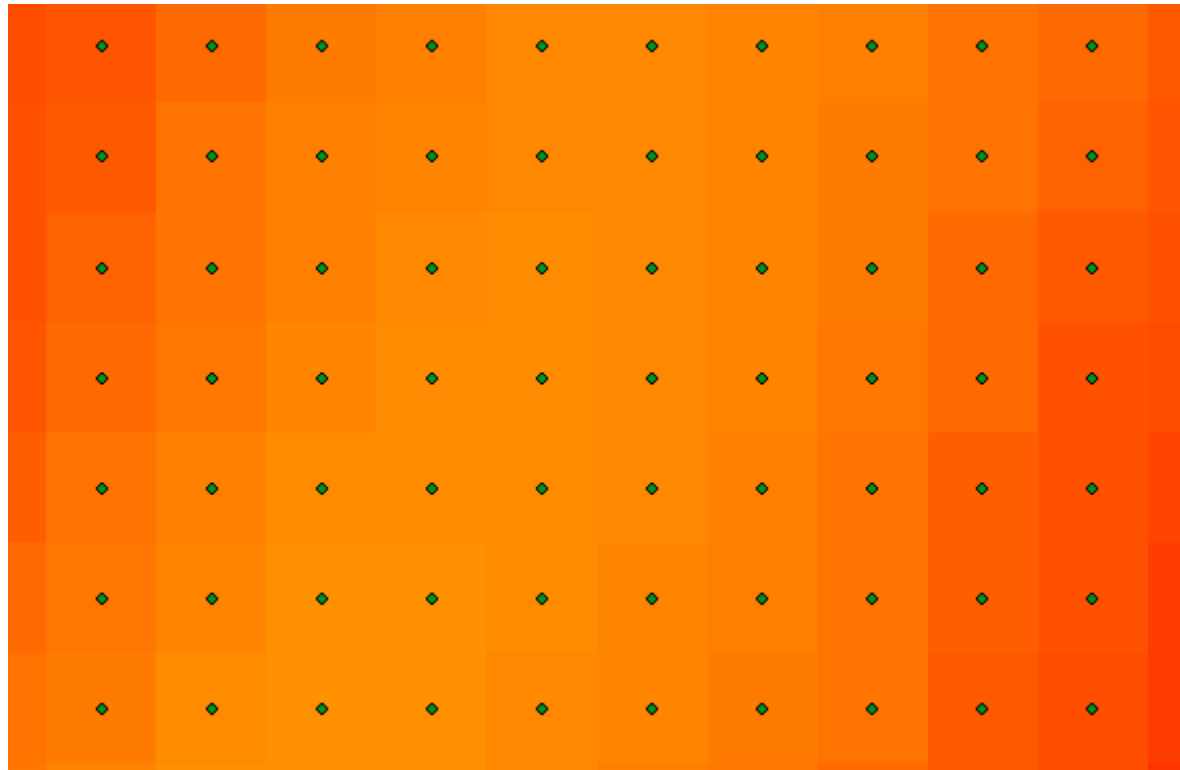


DEM

(Raster converted to Feature –points)



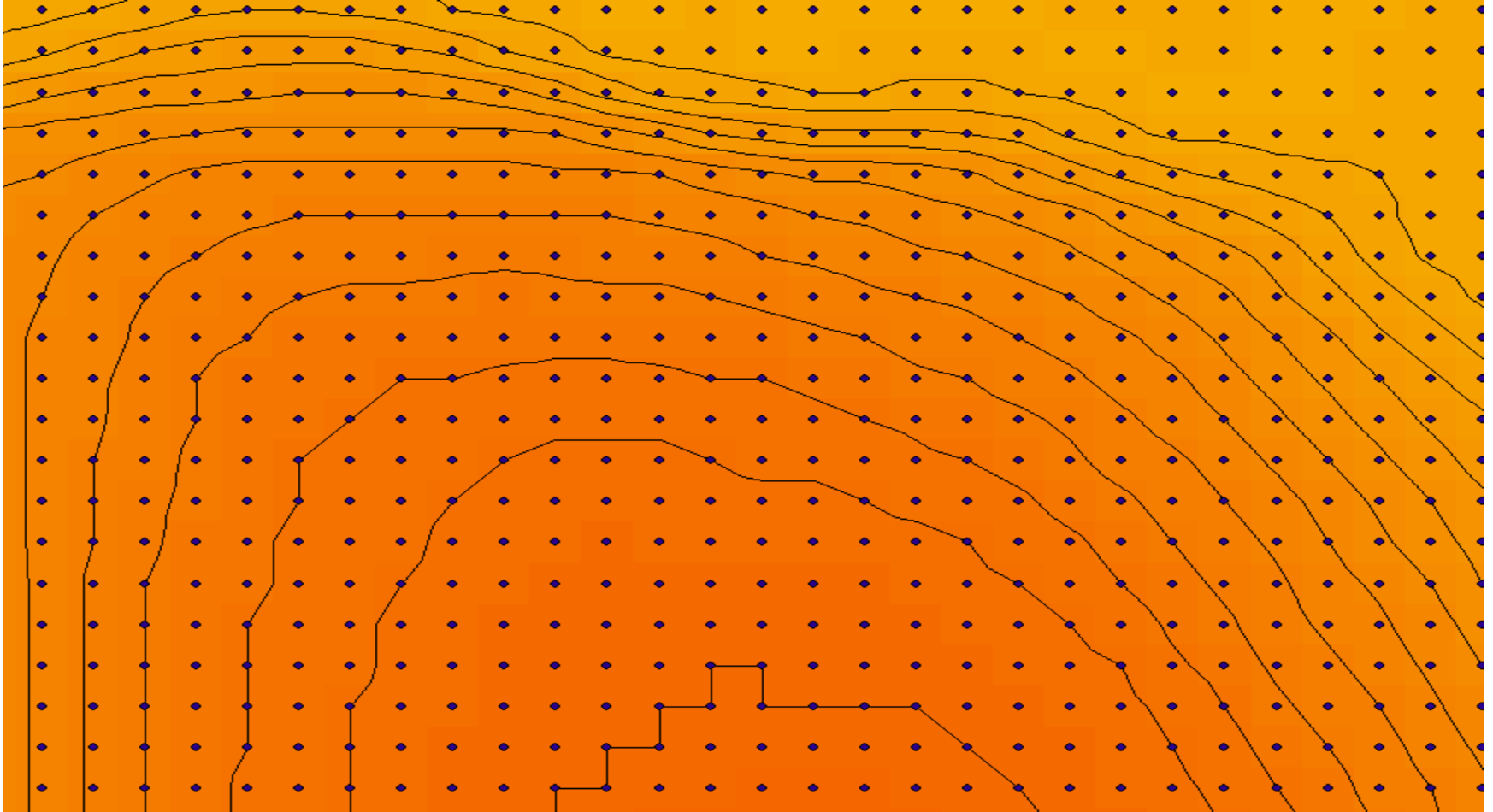
DEM Stores the elevation at the center of the cell



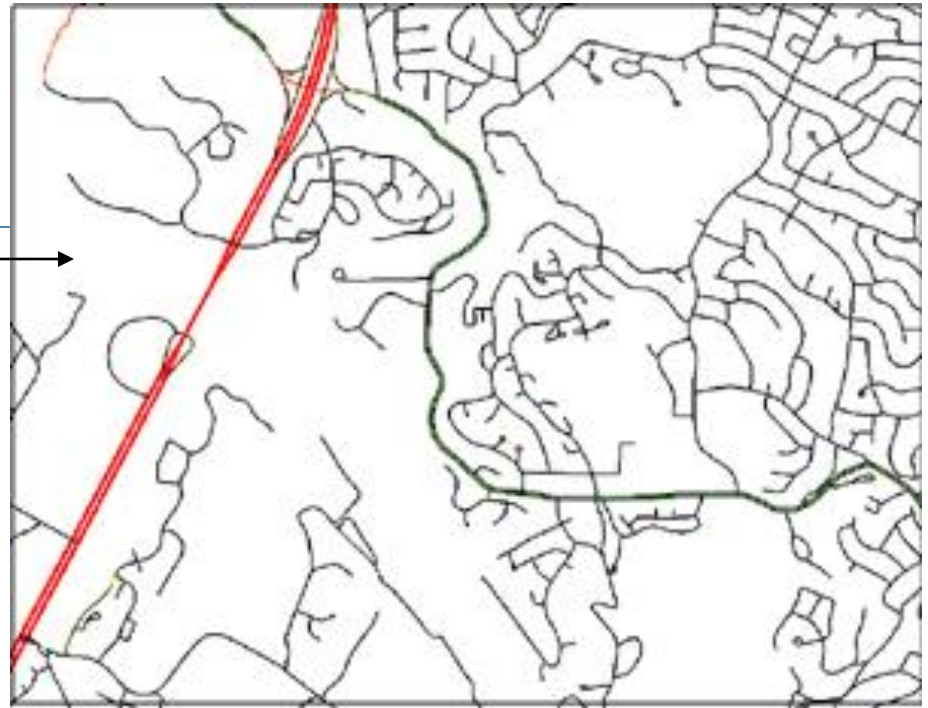
DEM with Elevations

955	952	949	944	940	939	937	937	938	939	942	945	948	950	953	957
954	951	948	943	939	938	937	937	938	940	943	946	949	952	954	957
952	950	946	942	939	937	936	937	938	940	944	948	951	953	956	958
951	949	945	941	938	936	936	937	938	941	945	950	952	955	957	960
949	947	943	939	936	936	936	937	939	942	947	951	953	956	958	961
947	944	941	938	935	935	936	938	939	942	947	951	954	957	960	962
945	943	940	936	934	935	937	938	940	943	948	952	955	958	961	963
943	941	938	935	934	935	937	939	941	944	948	952	956	959	962	964
942	940	937	935	934	935	937	939	942	945	949	953	957	960	963	965
940	938	936	934	933	934	937	940	943	946	950	954	958	961	963	966
939	937	935	934	933	934	937	941	944	948	951	955	959	962	964	966

Contours (drawn using Spatial Analyst)

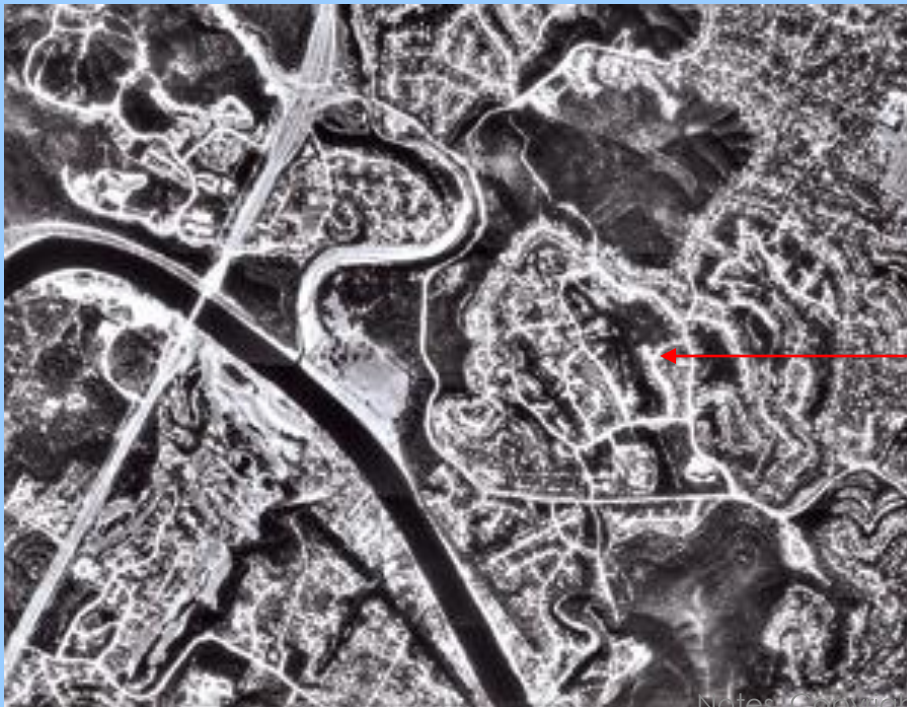


Digital Line Graph (DLG) of Road



Digital Line Graph (DLG) of River

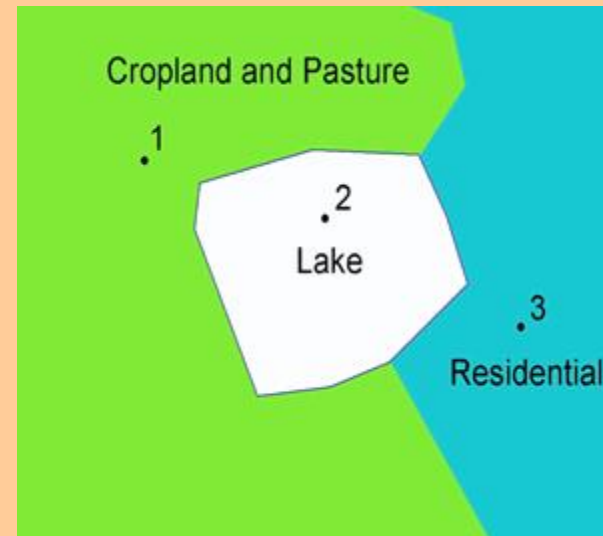
Scanned Topographic Map
Digital Raster Graphic (DRG)



Digital orthophoto quadrangle (DOQ).

Raster Representation of real-world information

1	1	1	1	1	1	1	3	3	3
1	1	1	1	1	1	1	3	3	3
1	1	1	1	1	1	3	3	3	3
1	1	1	2	2	2	2	3	3	3
1	1	1	2	2	2	2	3	3	3
1	1	1	2	2	2	2	3	3	3
1	1	1	1	2	2	2	3	3	3
1	1	1	1	1	1	3	3	3	3
1	1	1	1	1	1	1	3	3	3
1	1	1	1	1	1	1	1	3	3

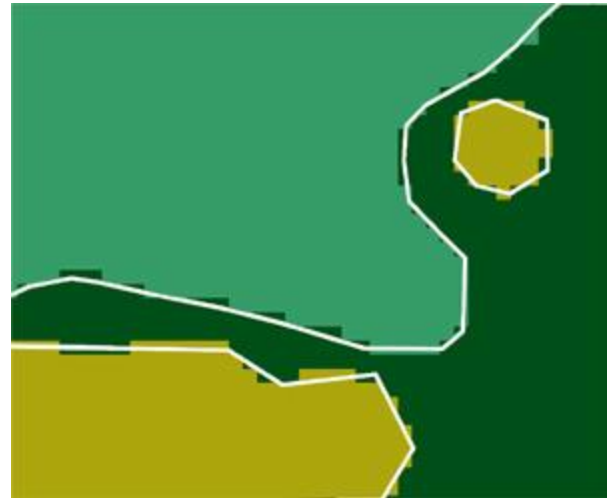


Raster and Vector Representation (based on raster) of real world

Raster



Vector



Spatial Data Formats

- CAD Drawings (Vector)
 - DXF (digital exchange format) Format
 - DWG Format
- ARC/INFO Coverage File (Vector)
- ARC/INFO Grid File (Raster)
- ARCVIEW Shapefile (Vector)

GIS Image Data Formats

- GIF
- TIFF
- JPEG
- ERDAS (ESRI data format)

Database Formats

- INFO Tables (ARC/INFO internal format)
- dBase Files
- Oracle Files
- Comma delimited text files
- Tables exported from EXCEL, ACCESS

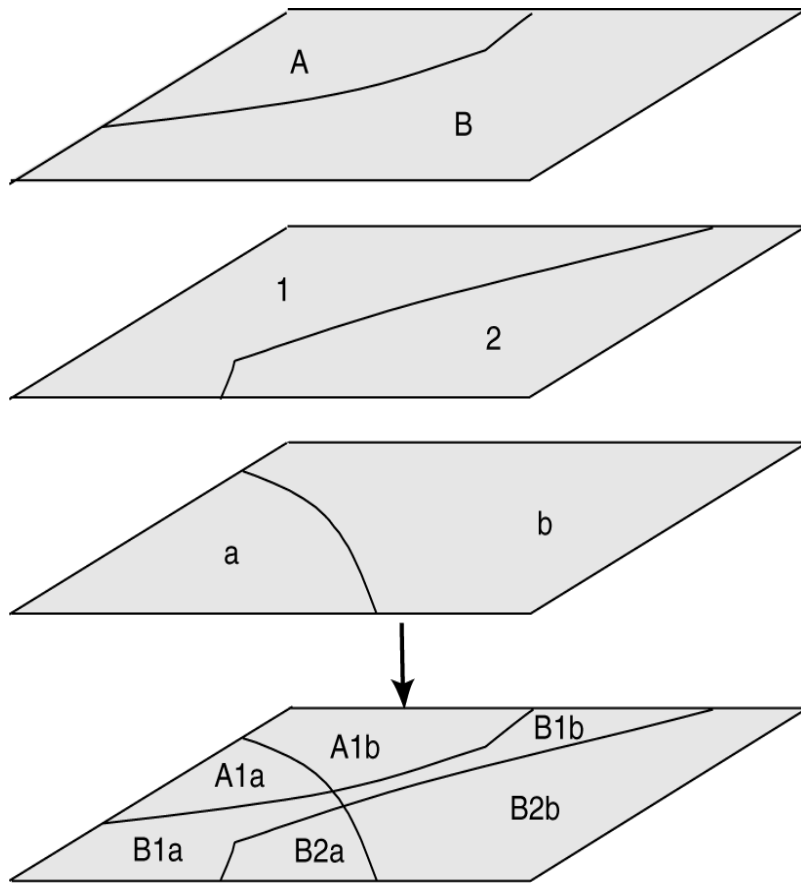
Classification of Data Formats (ESRI)

	Geo-relational	Object-based
Topological	Coverage	Geodatabase
Non-topological	Shapefile	Geodatabase

GIS operations

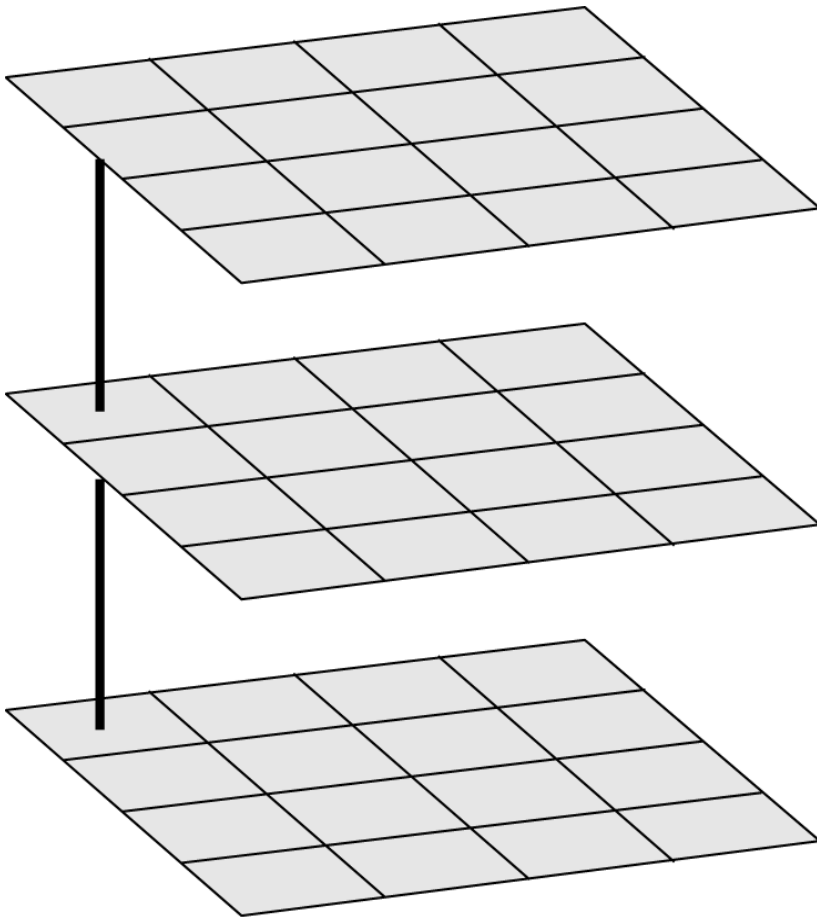
Spatial data input	<ol style="list-style-type: none">1. Data entry: use existing data, create new data2. Data editing3. Geometric transformation4. Projection and reprojection
Attribute data management	<ol style="list-style-type: none">1. Data entry and verification2. Database management3. Attribute data manipulation
Data display	<ol style="list-style-type: none">1. Cartographic symbolization2. Map design
Data exploration	<ol style="list-style-type: none">1. Attribute data query2. Spatial data query3. Geographic visualization
Data analysis	<ol style="list-style-type: none">1. Vector data analysis: buffering, overlay, distance measurement, spatial statistics, map manipulation2. Raster data analysis: local, neighborhood, zonal, global, raster data manipulation3. Terrain mapping and analysis4. Viewshed and watershed5. Spatial interpolation6. Geocoding and dynamic segmentation7. Path analysis and network applications
GIS modeling	<ol style="list-style-type: none">1. Binary models2. Index models3. Regression models4. Process models

Vector Based Overlay



A vector-based overlay operation combines spatial data and attribute data from different layers to create the output.

Raster based Overlay



A raster data operation with multiple rasters can take advantage of the fixed cell locations.

GIS Applications/Links/Resources

- U.S. Department of Labor: emerging fields
- <http://www.careervoyages.gov/>
- U.S. Geological Survey National Map
- <http://nationalmap.usgs.gov>
- U.S. Census Bureau On-Line Mapping Resources
- <http://www.census.gov/geo/www/maps/>
- U.S. Department of Housing and Urban Development
- <http://www.hud.gov/offices/cio/emap/index.cfm>
- U.S. Department of Health and Human Services
- <http://datawarehouse.hrsa.gov/>
- Department of Homeland Security's National Incident Management System
- <http://www.dhs.gov/>
- National Institute of Justice: crime mapping
- <http://www.ojp.usdoj.gov/nij/maps/>
- Federal Emergency Management Agency: flood insurance rate map
- http://www.fema.gov/plan/prevent/fhm/mm_main.shtm
- Larimer County, Colorado: land records
- <http://www.larimer.org/>
- Microsoft
- <http://www.microsoft.com/>
- Oracle
- <http://www.oracle.com/>

Links/Resources

- IBM: Spatial DataBlade
- <http://www-306.ibm.com/software/data/informix/blades/spatial/>
- National Association of Realtors
- <http://www.realtor.com/>
- Dodgeball
- <http://www.dodgeball.com/>
- Eyebeam: Fundrace
- <http://www.fundrace.org>
- Murals of Winnipeg
- <http://www.themuralsofwinnipeg.com/index.php>
- Arkansas Game and Fish Commission
- <http://vestig.cast.uark.edu/website/waterfowl/>
- ESRI
- <http://www.esri.com/>
- Autodesk
- <http://www3.autodesk.com/>
- Land Management Information Center at Minnesota Planning: EPPL7
- <http://www.lmic.state.mn.us/EPPL7/>
- Baylor University: GRASS
- <http://grass.baylor.edu/>

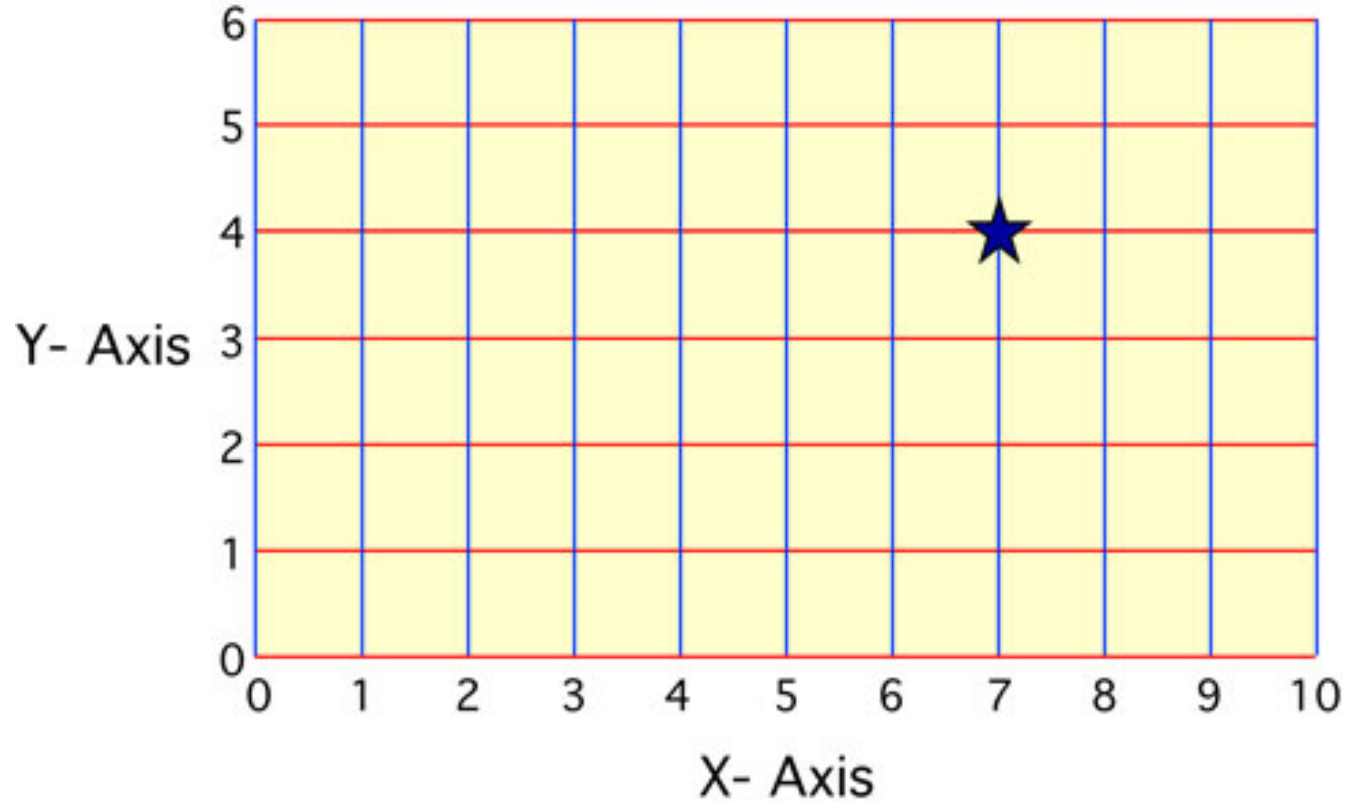
Links/Resources

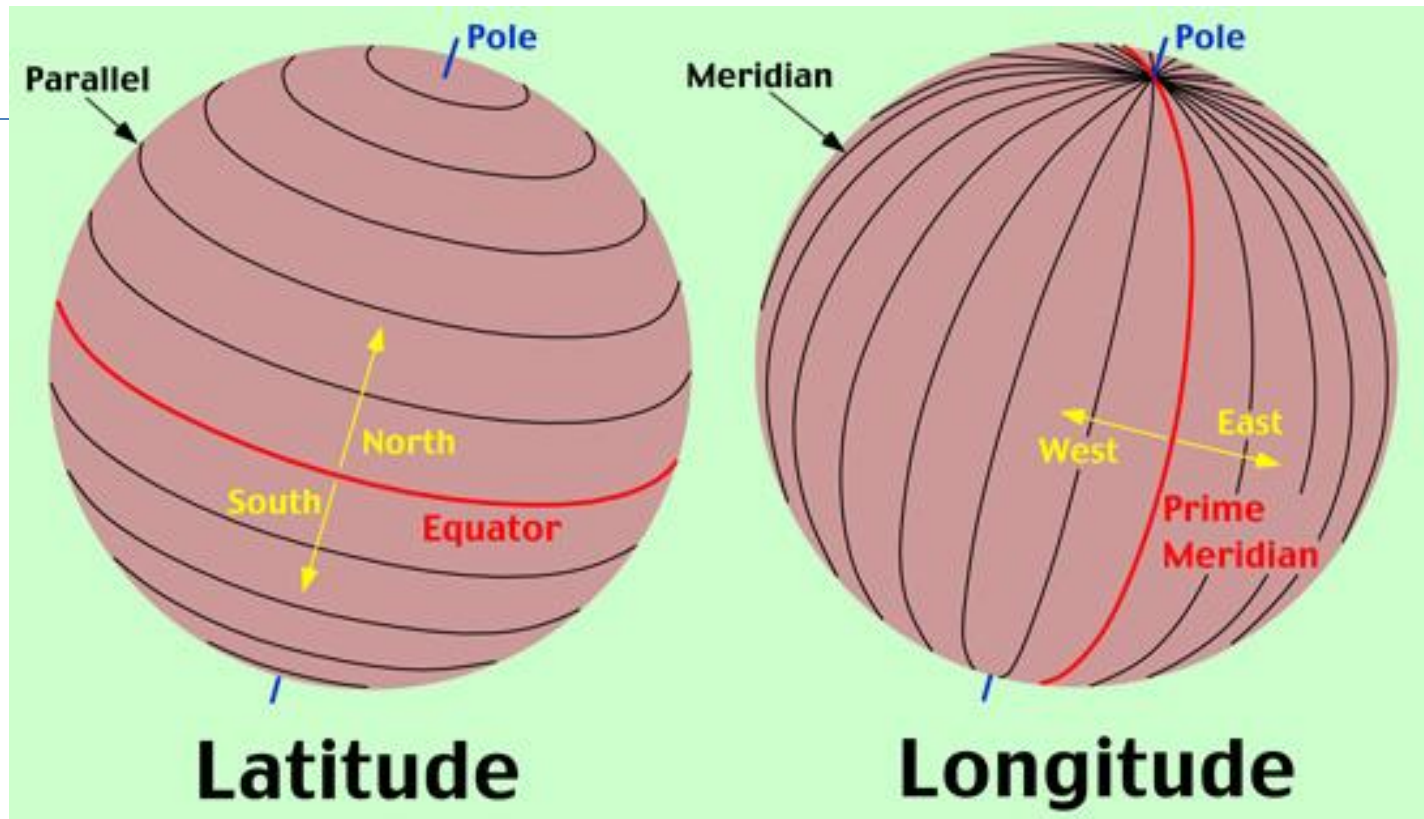
- Clark Labs: IDRISI
- <http://www.clarklabs.org/>
- International Institute for Aerospace Survey and Earth Sciences, the Netherlands: ILWIS
- <http://www.itc.nl/ilwis/>
- Manifold.net
- <http://www.manifold.net/>
- MapInfo Corporation
- <http://www.mapinfo.com/>
- Keigan Systems: MFworks
- <http://www.keigansystems.com/>
- Intergraph Corporation: MGE, GeoMedia
- <http://www.intergraph.com/>
- Bentley Systems, Inc: Microstation
- <http://www2.bentley.com/>
- PCI Geomatics: Geomatica
- <http://www.pcigeomatics.com/>
- Caliper Corporation: TransCAD, Maptitude
- <http://www.caliper.com/>

Spatial

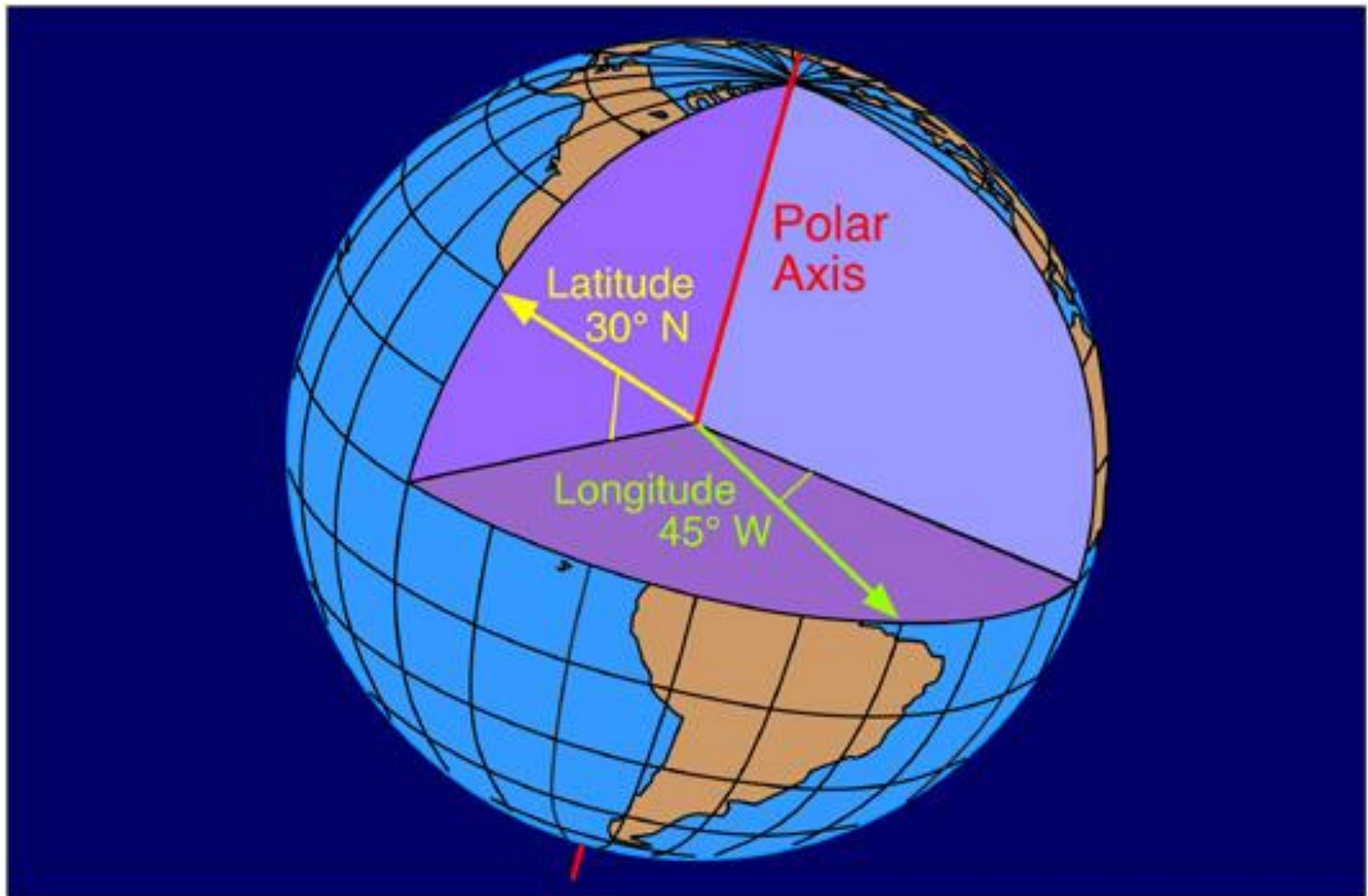
- Coordinate Systems

Cartesian Coordinate System





Spherical Grid System



Spatial Coordinate Systems

- Working in a GIS system, we need to be able to go from one coordinate system to the other:
 - Decimal Degree Approach
 - (DEGREES, MINUTES, SECONDS converted to
 - Decimal degrees as =
 - $\text{Degrees} + (\text{minutes}/60) + (\text{seconds}/3600)$
 - Used by ArcGIS
 - Projection Approach

Geodesy and Map Projections

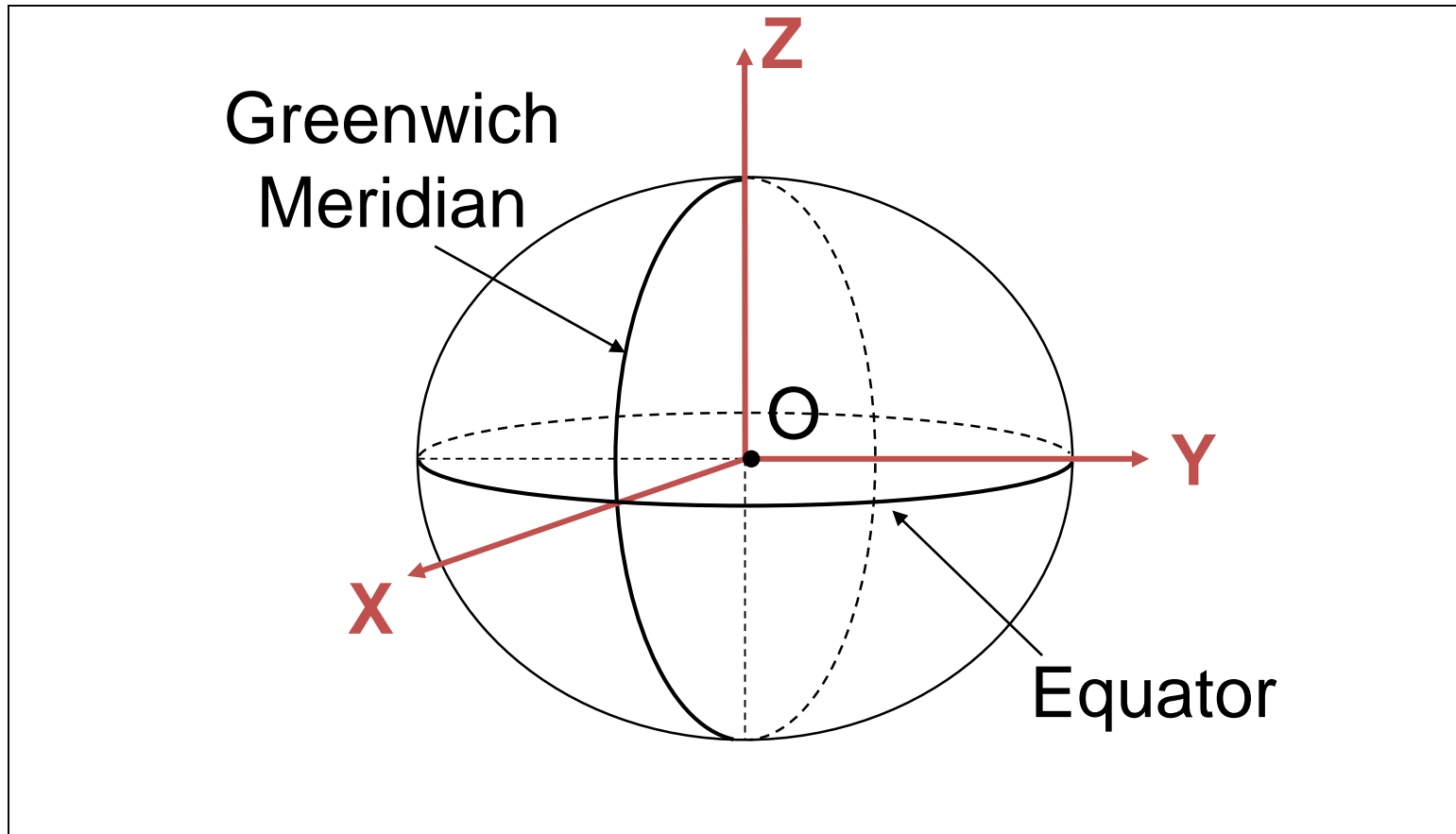
- **Geodesy** - the shape of the earth and definition of earth datum
- **Map Projection** - the transformation of a curved earth to a flat map
- **Coordinate systems** - (x,y) coordinate systems for map data

-
- (1) Global Cartesian coordinates (x, y, z) for the whole earth
 - (2) Geographic coordinates (ϕ, λ, z)
 - (3) Projected coordinates (x, y, z) on a local area of the earth's surface
 - The z -coordinate in (1) and (3) is defined geometrically; in (2) the z -coordinate is defined gravitationally

-
- The geographic coordinate system is the location reference system for spatial features on the Earth's surface.
 - The geographic coordinate system is defined by longitude and latitude.

-
- Geographic coordinates are the earth's latitude and longitude system, ranging from 90 degrees south to 90 degrees north in latitude and 180 degrees west to 180 degrees east in longitude.
 - A line with a constant latitude running east to west is called a parallel.
 - A line with constant longitude running from the north pole to the south pole is called a meridian.
 - The zero-longitude meridian is called the prime meridian and passes through Greenwich, England.
 - A grid of parallels and meridians shown as lines on a map is called a graticule.

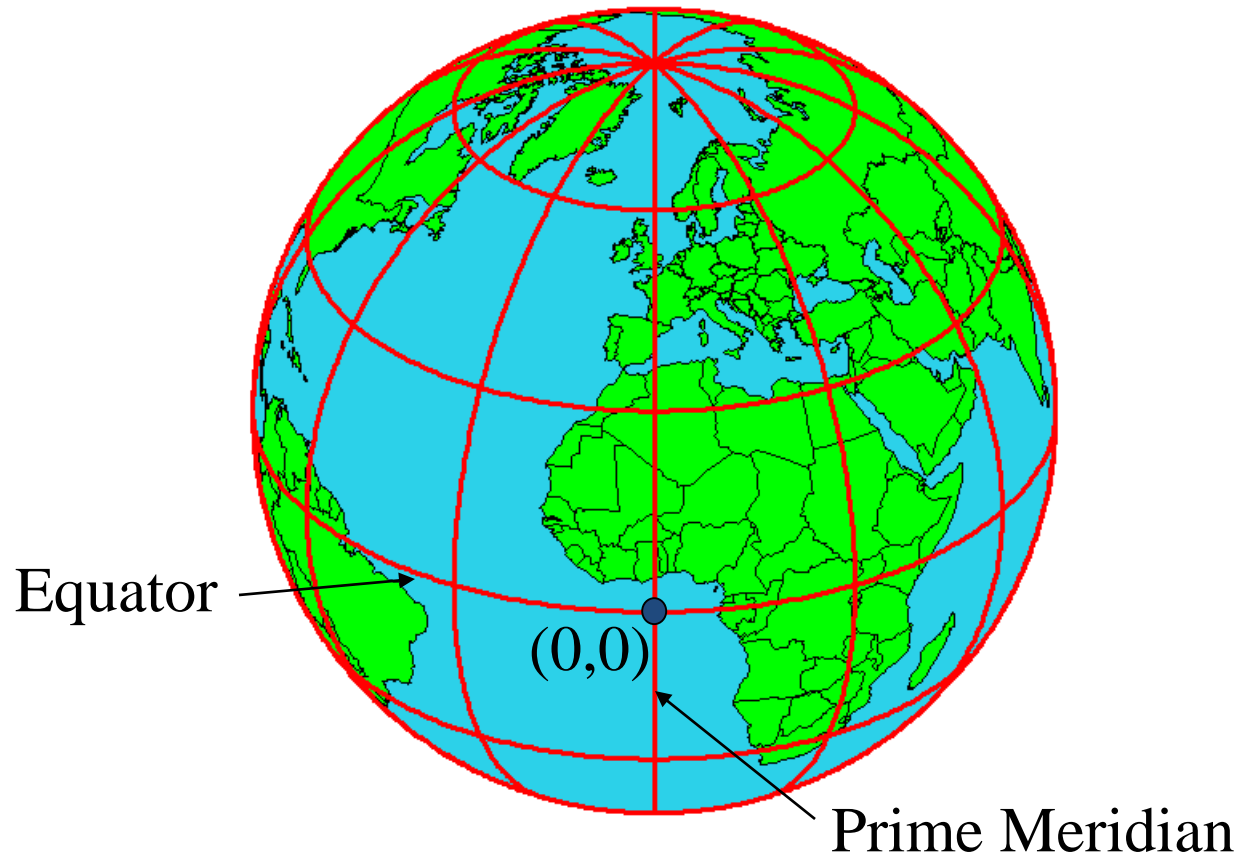
Global Cartesian Coordinates (x,y,z)



Geographic Coordinates (ϕ , λ , z)

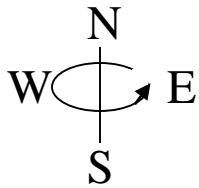
- Latitude (ϕ) and Longitude (λ) defined using an **ellipsoid**, an ellipse rotated about an axis
- Elevation (z) defined using **geoid**, a surface of constant gravitational potential
- Earth **datums** define standard values of the ellipsoid and geoid

Origin of Geographic Coordinates



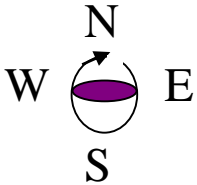
Latitude and Longitude

Longitude line (Meridian)

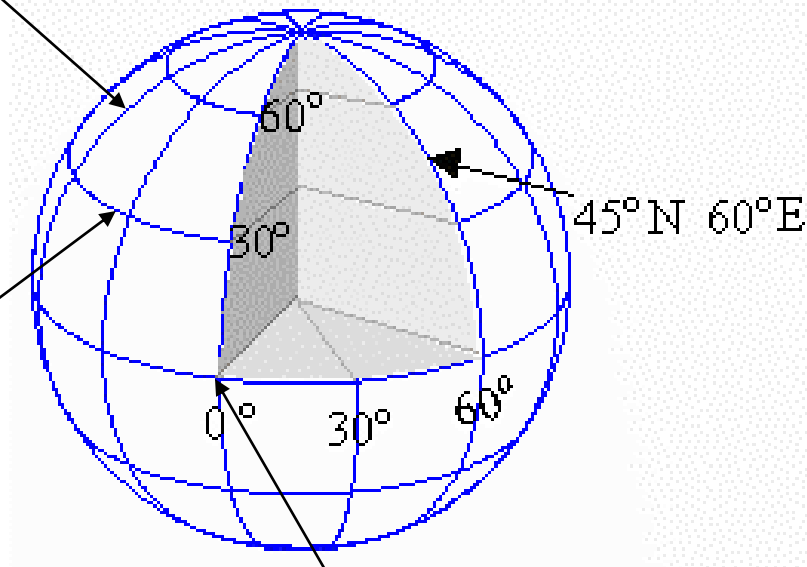


Range: 180°W - 0° - 180°E

Latitude line (Parallel)



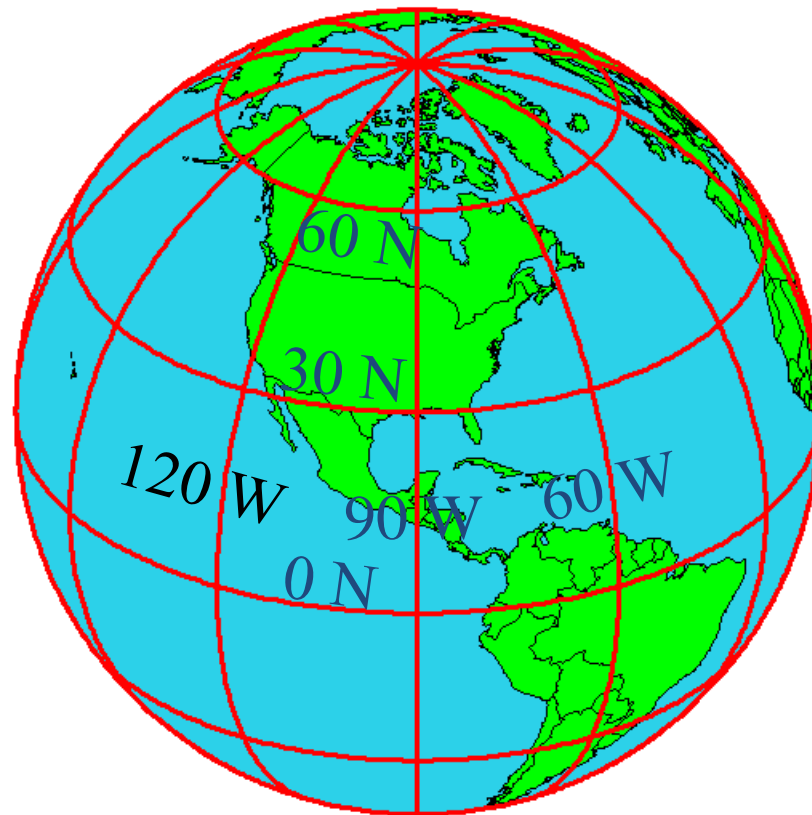
Range: 90°S - 0° - 90°N



$(0^{\circ}\text{N}, 0^{\circ}\text{E})$

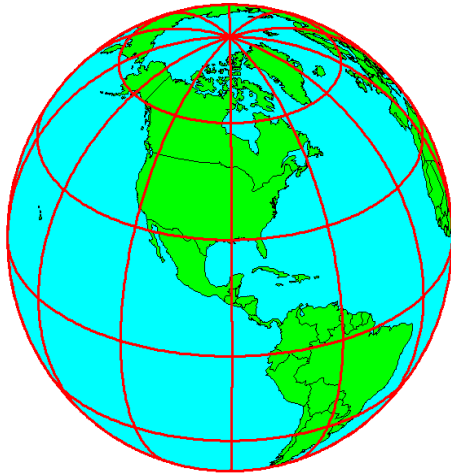
Equator, Prime Meridian

Latitude and Longitude in North America

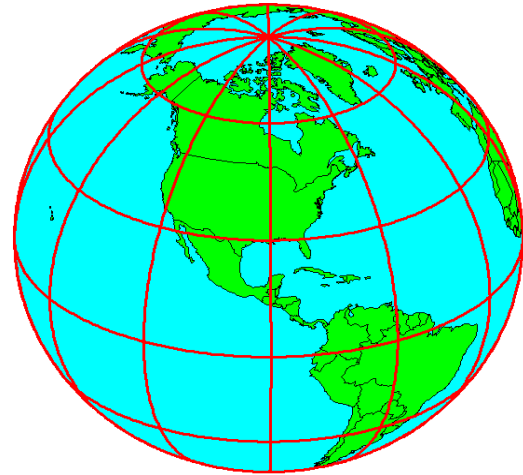


Shape of the Earth

We think of the earth as a **sphere**



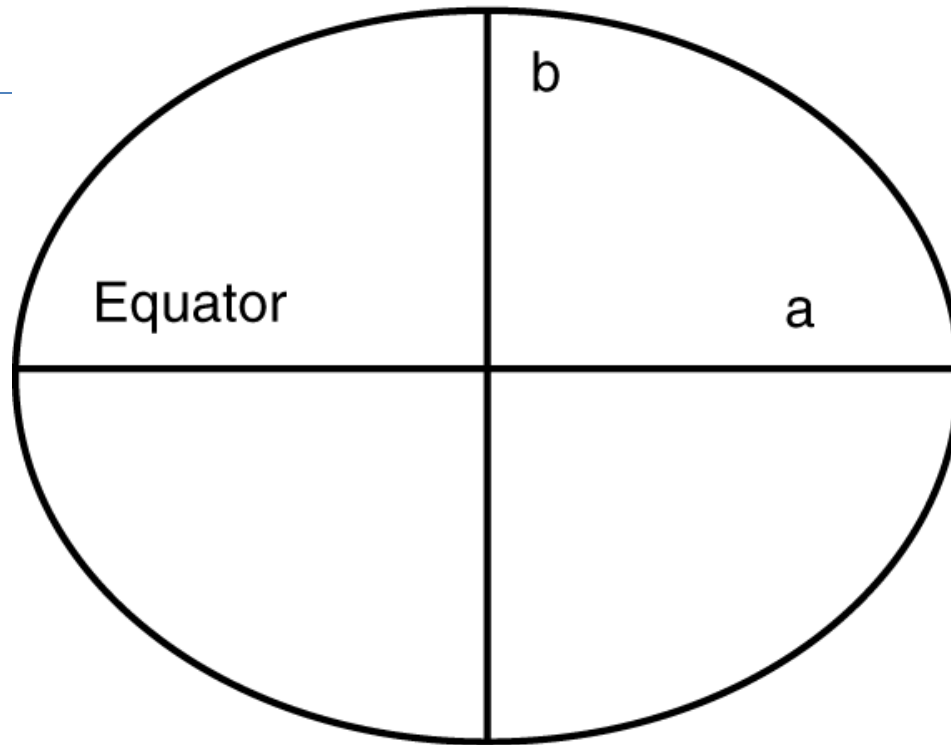
It is actually a **spheroid**, slightly larger in radius at the equator than at the poles



Approximation of Earth

- The simplest model is a sphere, which is typically used in discussing map projections.
- But the Earth is not a perfect sphere: the Earth is wider along the equator than between the poles.
- A better approximation to the shape of the Earth is a *spheroid*, also called *ellipsoid*, an ellipse rotated about its minor axis.

North Pole



South Pole

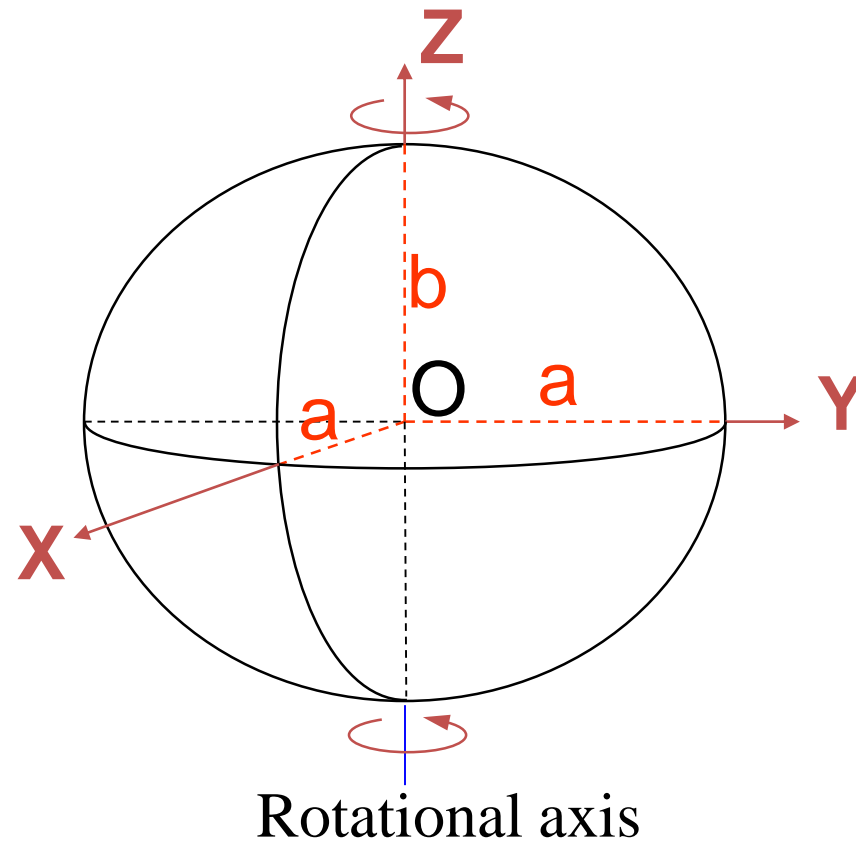
The flattening is based on the difference between the semi-major axis a and the semi-minor axis b .

Geographic Coordinates (ϕ , λ , z)

- Latitude (ϕ) and Longitude (λ) defined using an **ellipsoid**, an ellipse rotated about an axis
- Elevation (z) defined using **geoid**, a surface of constant gravitational potential
- Earth **datums** define standard values of the ellipsoid and geoid

Ellipsoid or Spheroid

Rotate an ellipse around an axis



-
- A *datum* is a mathematical model of the Earth, which serves as the reference or base for calculating the geographic coordinates of a location.
 - A shift of the datum will result in the shift of positions of points.

- A *datum* is a mathematical model of the Earth, which serves as the reference or base for calculating the geographic coordinates of a location.
- A shift of the datum will result in the shift of positions of points.

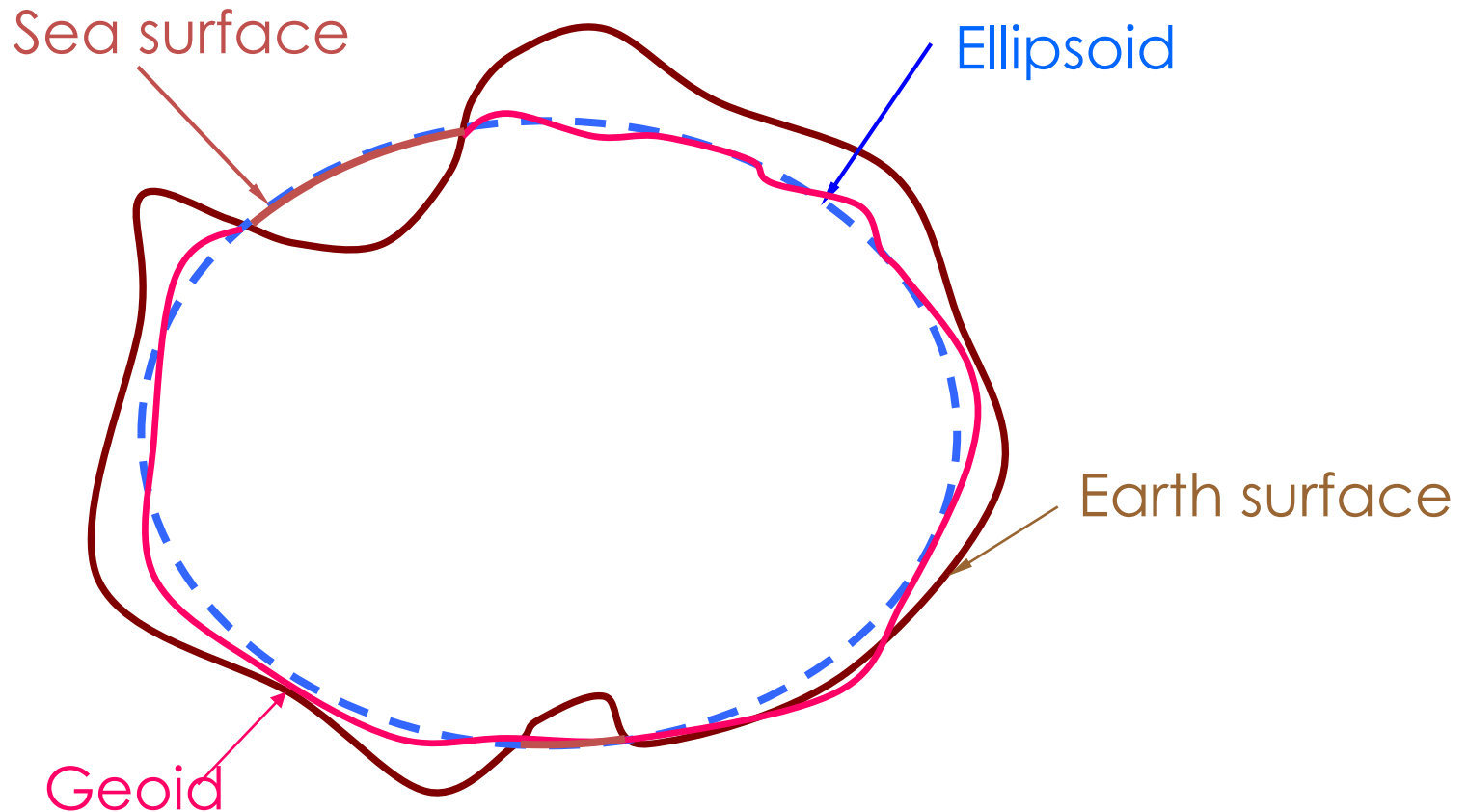
Datum

Horizontal Earth Datums

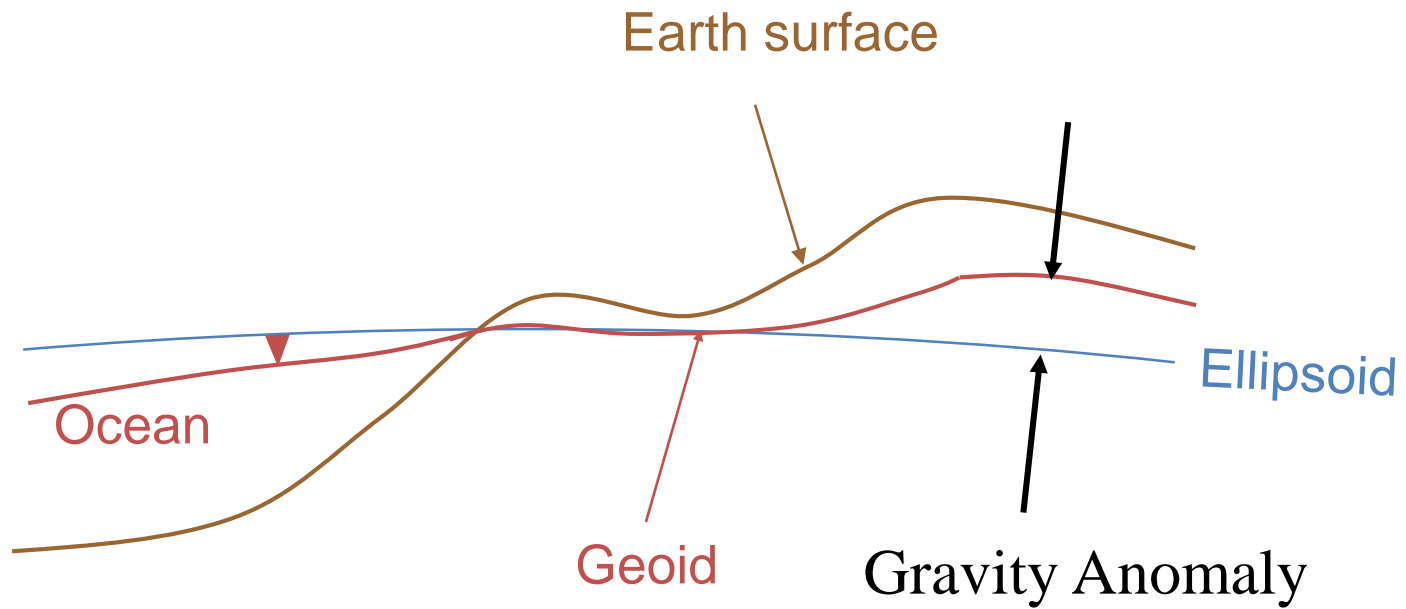
- An earth datum is defined by an **ellipse** and an **axis of rotation**
- **NAD27** (North American Datum of 1927) uses the Clarke (1866) ellipsoid on a non geocentric axis of rotation
- **NAD83** (NAD,1983) uses the GRS80 ellipsoid on a geocentric axis of rotation
- **WGS84** (World Geodetic System of 1984) uses GRS80, almost the same as NAD83

Representations of the Earth

Mean Sea Level is a surface of constant gravitational potential called the **Geoid**



Geoid and Ellipsoid



Vertical Earth Datums

- A vertical datum defines elevation, z
- **NGVD29** (National Geodetic Vertical Datum of 1929)
- **NAVD88** (North American Vertical Datum of 1988)
- Takes into account a map of gravity anomalies between the ellipsoid and the geoid

Definitions

- Geodesy - the shape of the earth and definition of earth datums
- Map Projection - the transformation of a curved earth to a flat map
- Coordinate systems - (x,y)
coordinate systems for map data

-
- A map projection is a systematic arrangement of parallels and meridians on a plane surface.
 - Cartographers group map projections by the preserved property into conformal, equal area or equivalent, equidistant, and azimuthal or true direction.
 - Cartographers also use a geometric object (a cylinder, cone, or plane) and a globe (i.e., a sphere) to illustrate how to construct a map projection.

Map Projection



Curved Earth

**Geographic coordinates: f, l
(Latitude & Longitude)**



Flat Map

**Cartesian coordinates:
 x, y
(Easting & Northing)**

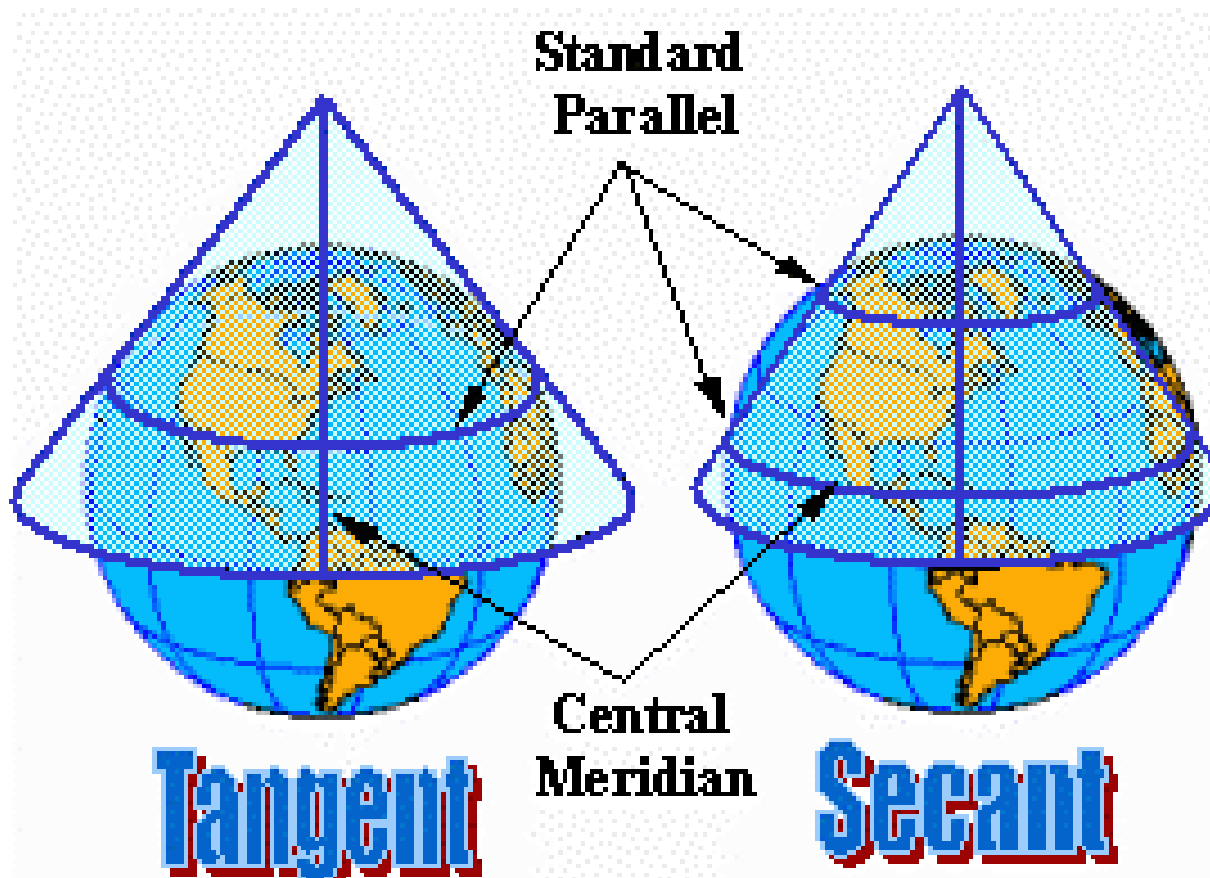


Types of Projections

- **Conic** (Albers Equal Area, Lambert Conformal Conic) - good for East-West land areas
- **Cylindrical** (Transverse Mercator) - good for North-South land areas
- **Azimuthal** (Lambert Azimuthal Equal Area) - good for global views

Conic Projections

(Albers, Lambert)

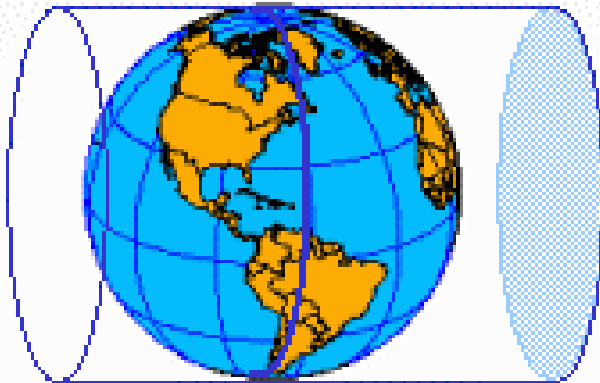
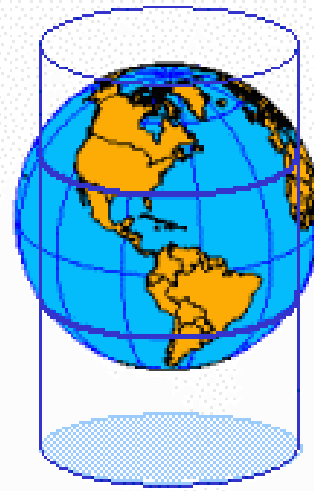


Cylindrical Projections

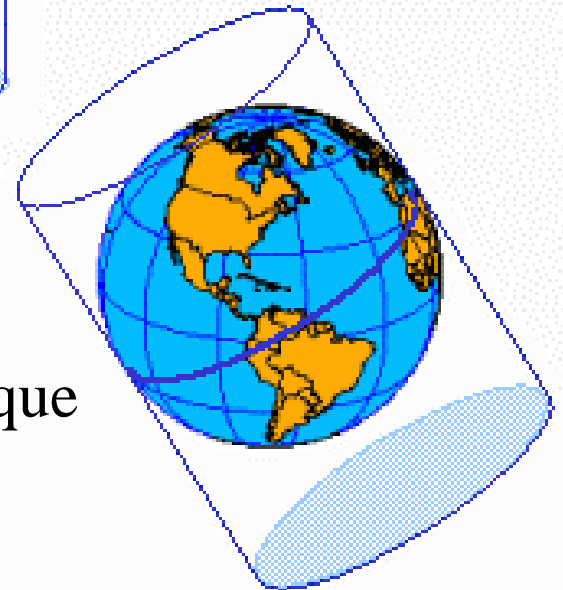
(Mercator)



Transverse



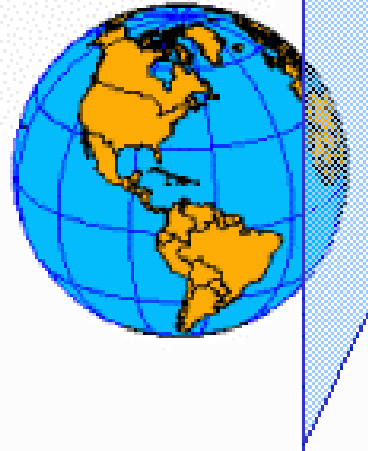
Oblique



Azimuthal (Lambert)



Polar

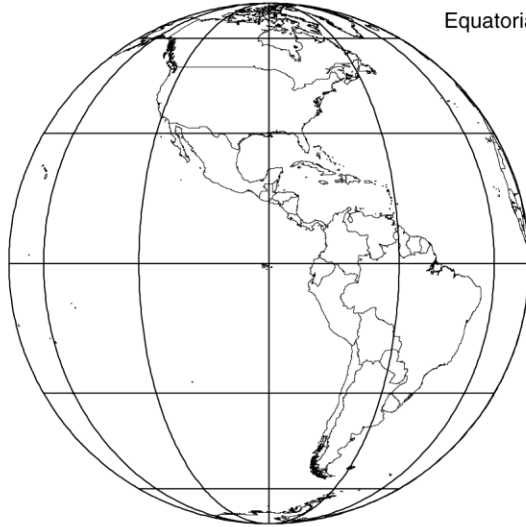


Equatorial



Oblique

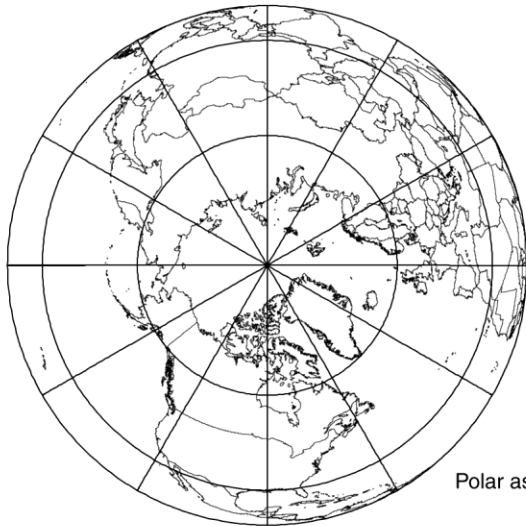
Equatorial aspect



Oblique aspect



Polar aspect



Projections Preserve Some Earth Properties

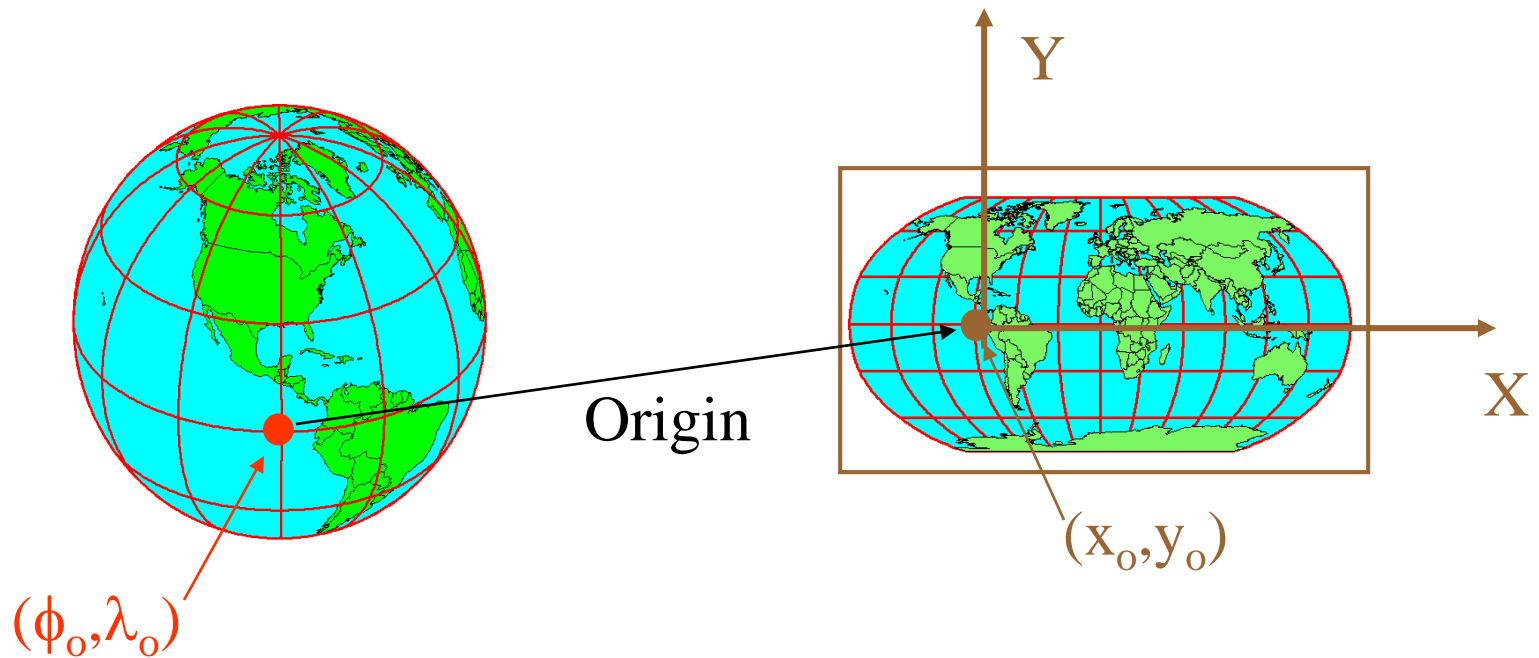
- **Area** - correct earth surface area (Albers Equal Area) important for mass balances
- **Shape** - local angles are shown correctly (Lambert **Conformal** Conic)
- **Direction** - all directions are shown correctly relative to the center (Lambert Azimuthal Equal Area)
- **Distance** - preserved along particular lines
- Some projections preserve **two** properties

Coordinate Systems

- Hydrologic calculations are done in **Cartesian** or **Planar coordinates** (x,y,z)
- Earth locations are measured in **Geographic coordinates** of latitude and longitude (ϕ,λ)
- **Map Projections** transform (ϕ,λ) (x,y)

Coordinate System

A planar coordinate system is defined by a pair of orthogonal (x,y) axes drawn through an origin



Standard Coordinate Systems

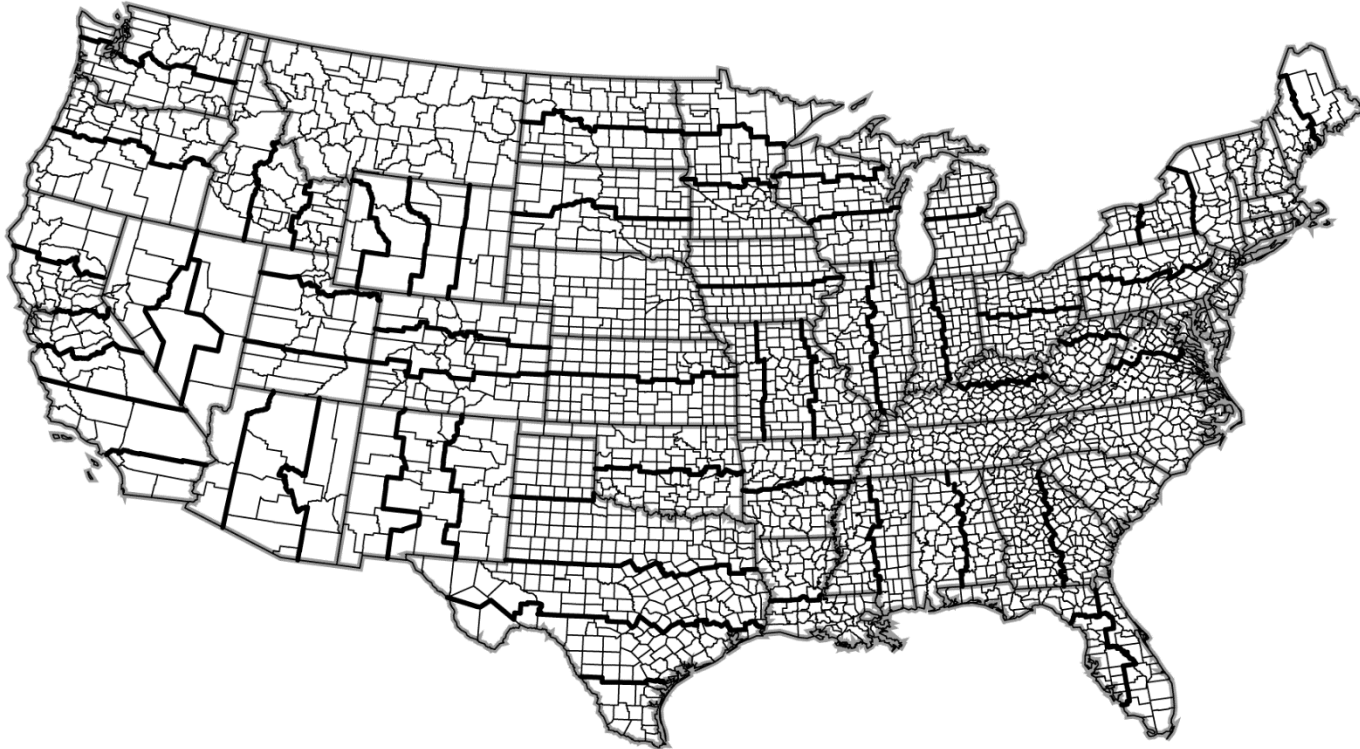
- Some standard coordinate systems used in the United States are
 - geographic coordinates (Un-projected data)
 - universal transverse Mercator system
 - military grid
 - state plane
- To compare or edge-match maps in a GIS, both maps **MUST** be in the same coordinate system.

-
- . Geographic Information referenced with Latitude and Longitude is
Un PROJECTED data

- Many States in the U.S. employ the
STATE PLANE COORDINATE (SPC) SYSTEM
- Many STATES have more than one SPC
corresponding to different parts of the state

e.g. Some states have two (south and north) plus
one (single plane) projection systems

State Plane Coordinate System



SPC83 zones in the conterminous United States. The thinner lines are county boundaries, and the gray lines are state boundaries.

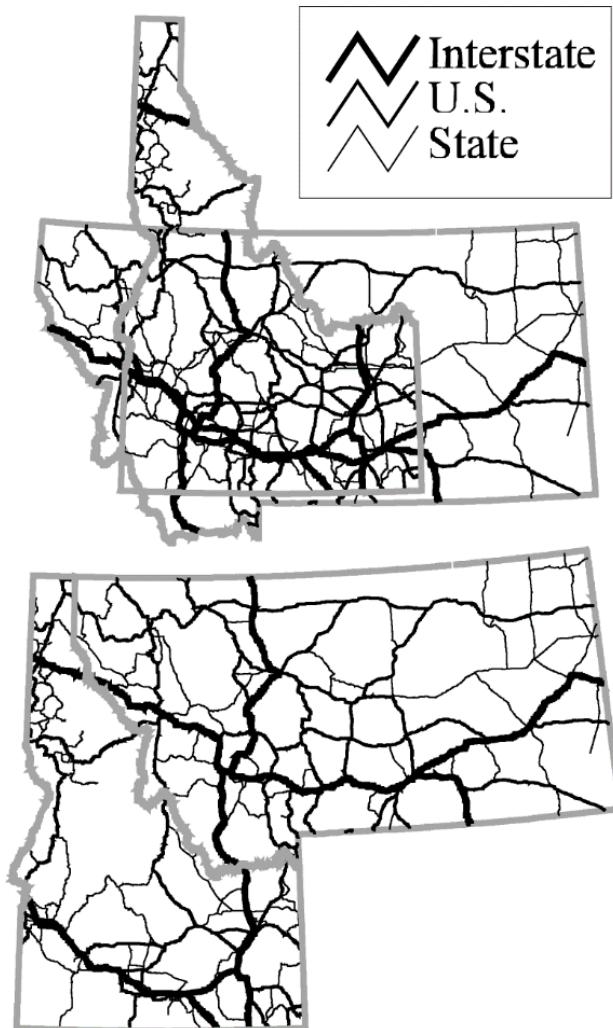
Universal Transverse Mercator Coordinate System



- Uses the **Transverse Mercator** projection
- Each zone has a **Central Meridian** (λ_o), zones are 6° wide, and go from pole to pole
- 60 zones cover the earth from East to West
- **Reference Latitude** (ϕ_o), is the equator
- (Xshift, Yshift) = false easting and northing so you never have a negative coordinate

Two map layers are not going to register spatially unless they are based on the same coordinate system.

Example of Incorrect Projections



The top map shows the road networks in Idaho and Montana based on different coordinate systems. The bottom map shows the road networks based on the same coordinate system.

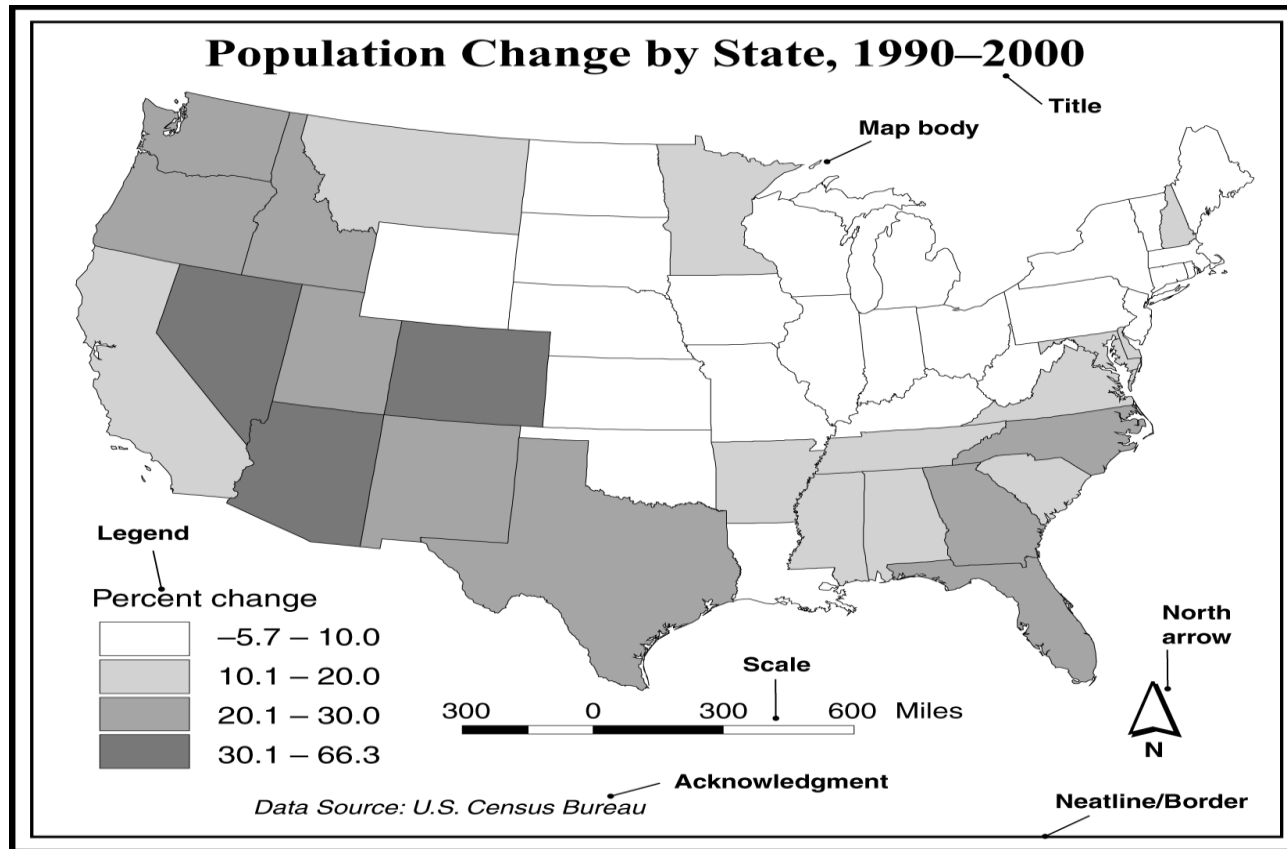
Acknowledgements

- Few Figures from this lecture notes are adopted from world wide web and other text-book based sources.
- Details of the figures and sources are available with the instructor.
- Some of the figures are adopted from text book of Kang-Tsang Chang.

Common Map Elements

- Common map elements are the title, body, legend, north arrow, scale, acknowledgment, and neatline/map border.
- Other elements include the graticule or grid, name of map projection, inset or location map, and data quality information.

Common map elements.



Spatial Features and Map Symbols

- To display a spatial feature on a map, we use a map symbol to indicate the feature's location and a visual variable, or visual variables, with the symbol to show the feature's attribute data.
- The general rule for vector data is to use point symbols for point features, line symbols for line features, and area symbols for area features.
- Visual variables for data display include hue, value, chroma, size, texture, shape, and pattern.



This map uses area symbols to show watersheds, a line symbol for streams, and a point symbol for gage stations.

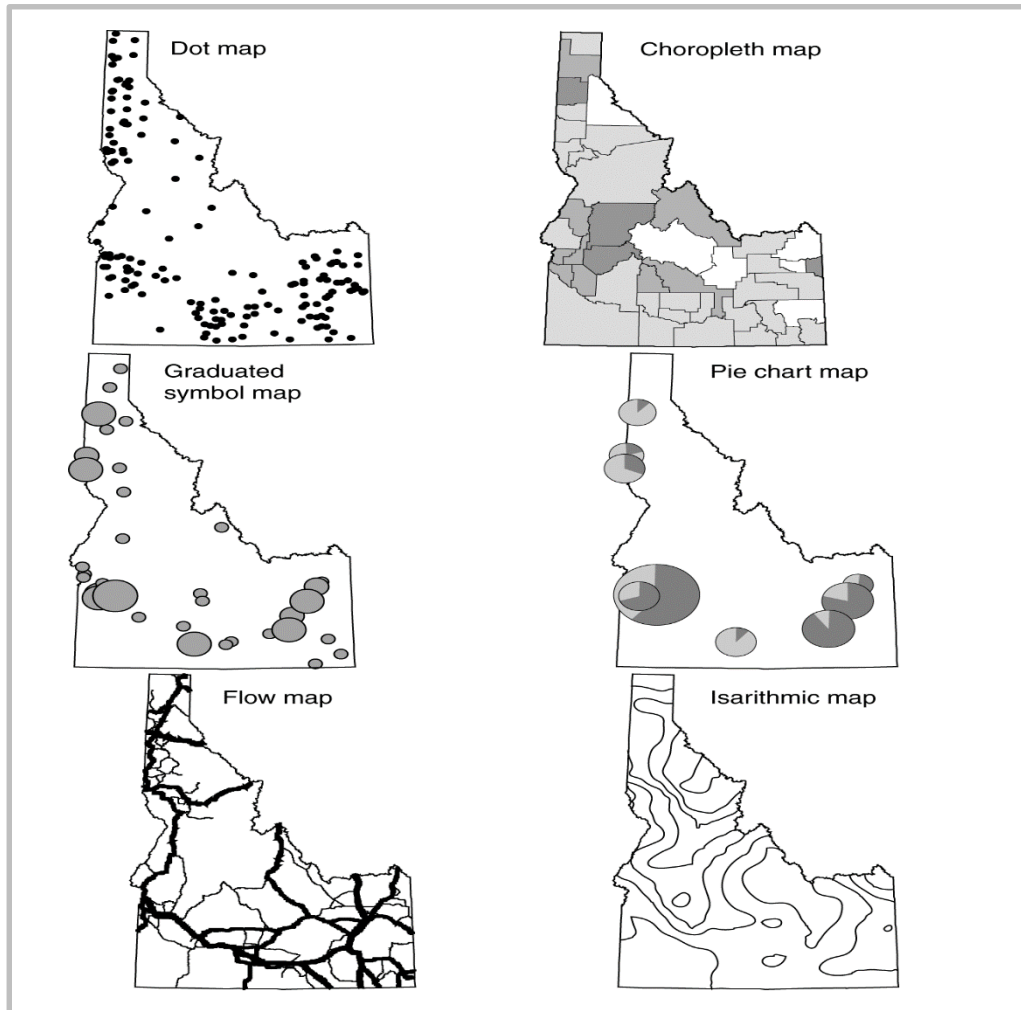
Data Classification

Five commonly used classification methods are: equal interval, equal frequency, mean and standard deviation, natural breaks, and user defined.

Types of Maps

- Maps can be general reference or thematic by function.
- Maps can be qualitative or quantitative by map symbol.
- Common types of quantitative maps include the dot map, the choropleth map, the dasymetric map, the graduated symbol map, the proportional symbol map, the chart map, the flow map, and the isarithmic map.

Quantitative Maps



Six common types of quantitative maps.

Typography

Text is needed for almost every map element. Mapmakers treat text as a map symbol because, like point, line, or area symbols, text can have many type variations.

Selection of Type Variations

- Cartographers recommend legibility, harmony, and conventions for selection of type variations.
- Mapmakers can generally achieve harmony by adopting only one or two typefaces on a map.

Too many Type faces



The look of the map is not harmonious because of too many typefaces.

Placement of Text

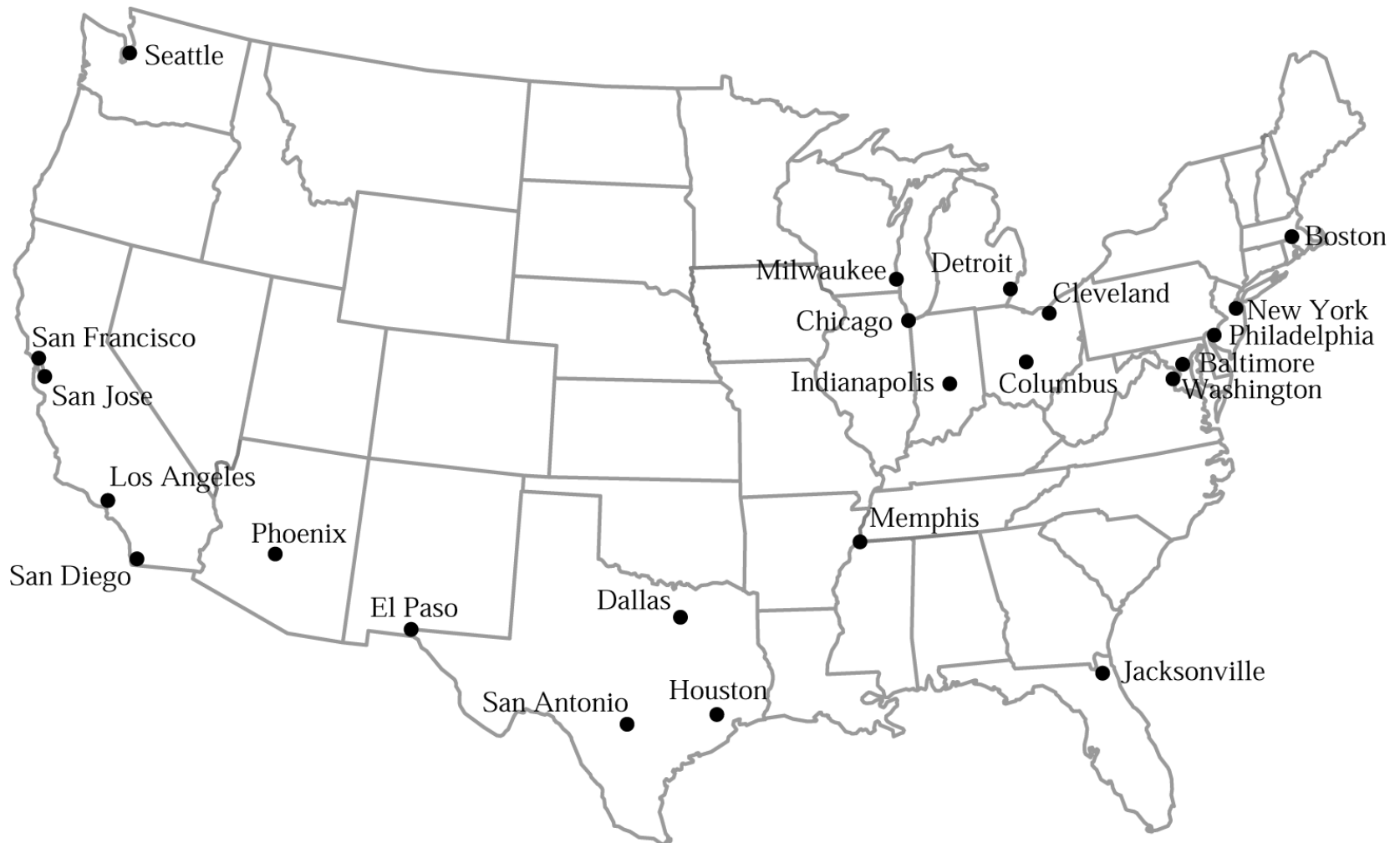
- As a general rule, a label should be placed to show the location or the area extent of the named spatial feature.
- ArcGIS offers interactive and dynamic labeling for placement of text in the map body.

Dynamic Labeling

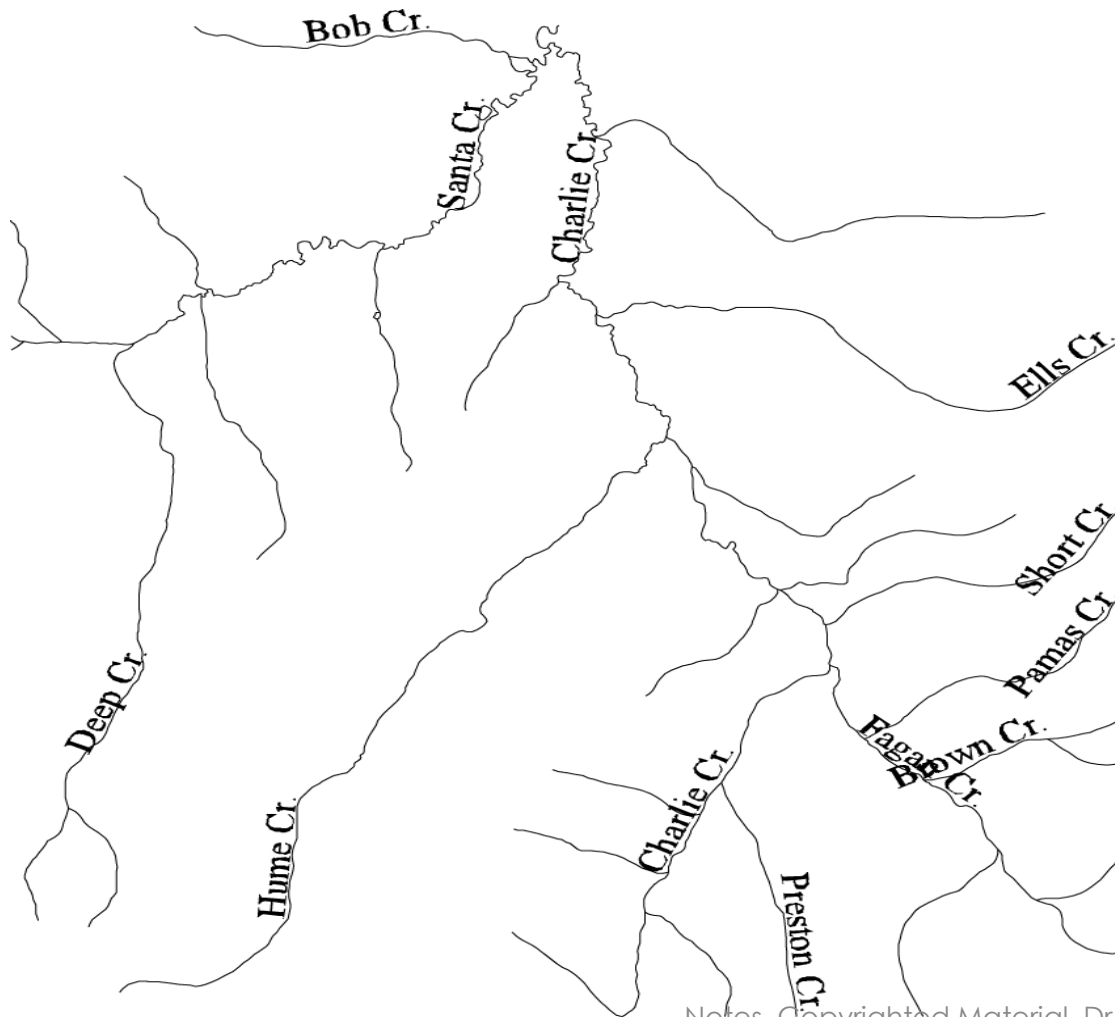


Dynamic labeling of major cities in the United States. The initial result is good but not totally satisfactory. Philadelphia is missing. Labels of San Antonio, Indianapolis, and Baltimore overlap slightly with point symbols. San Francisco is too close to San Jose.

Revised Labeling

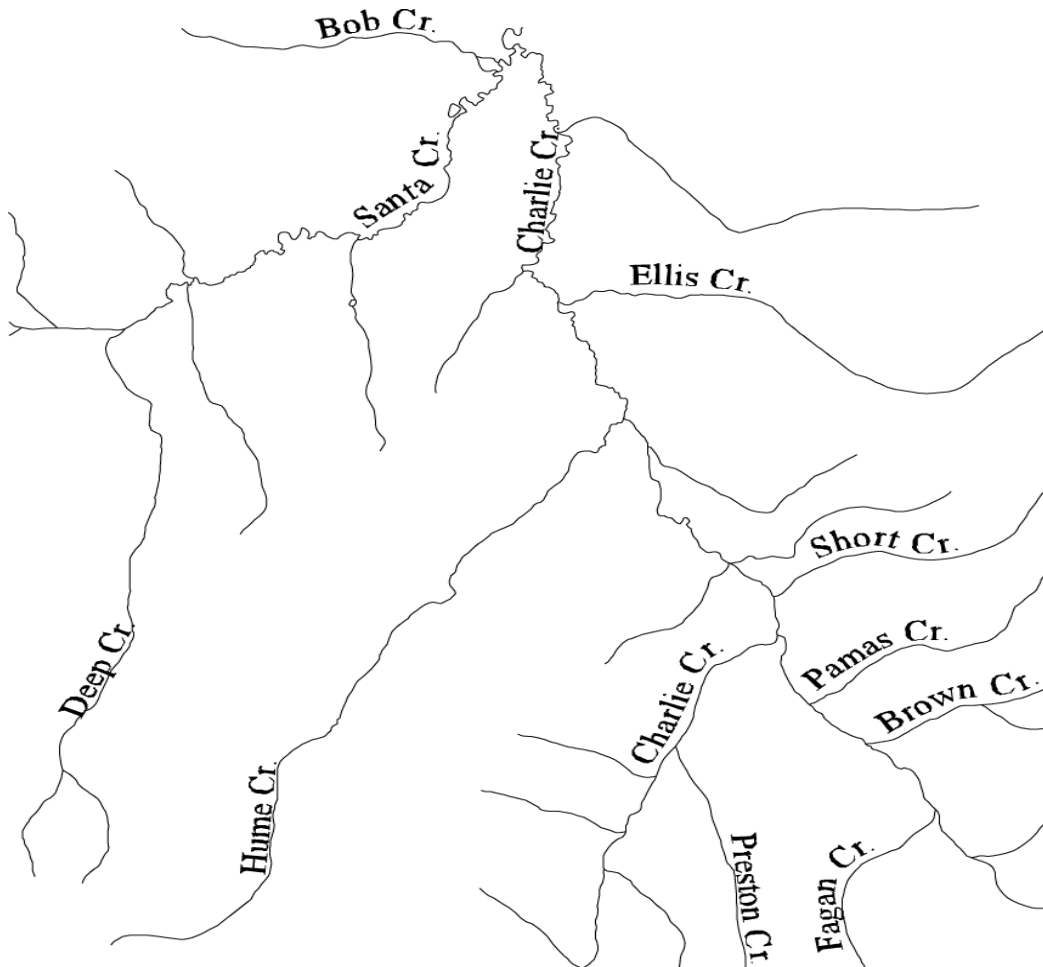


Dynamic Labeling



Dynamic labeling of streams may not work for every label. Brown Cr. overlaps with Fagan Cr., and Pamas Cr. and Short Cr. do not follow the course of the creek.

Spline Text Tool



Problem labels in Figure are redrawn with the spline text tool.

Map Design

- Map design is a visual plan to achieve a goal. A well-designed map is balanced, coherent, ordered, and interesting to look at, whereas a poorly designed map is confusing and disoriented. Map design is both an art and science.
- Cartographers usually study map design from the perspectives of layout and visual hierarchy.
- Layout deals with the arrangement and composition of various map elements on a map. Major concerns with layout are focus, order, and balance.
- Visual hierarchy is the process of developing a visual plan to introduce the 3-D effect or depth to maps.

Box- Maps

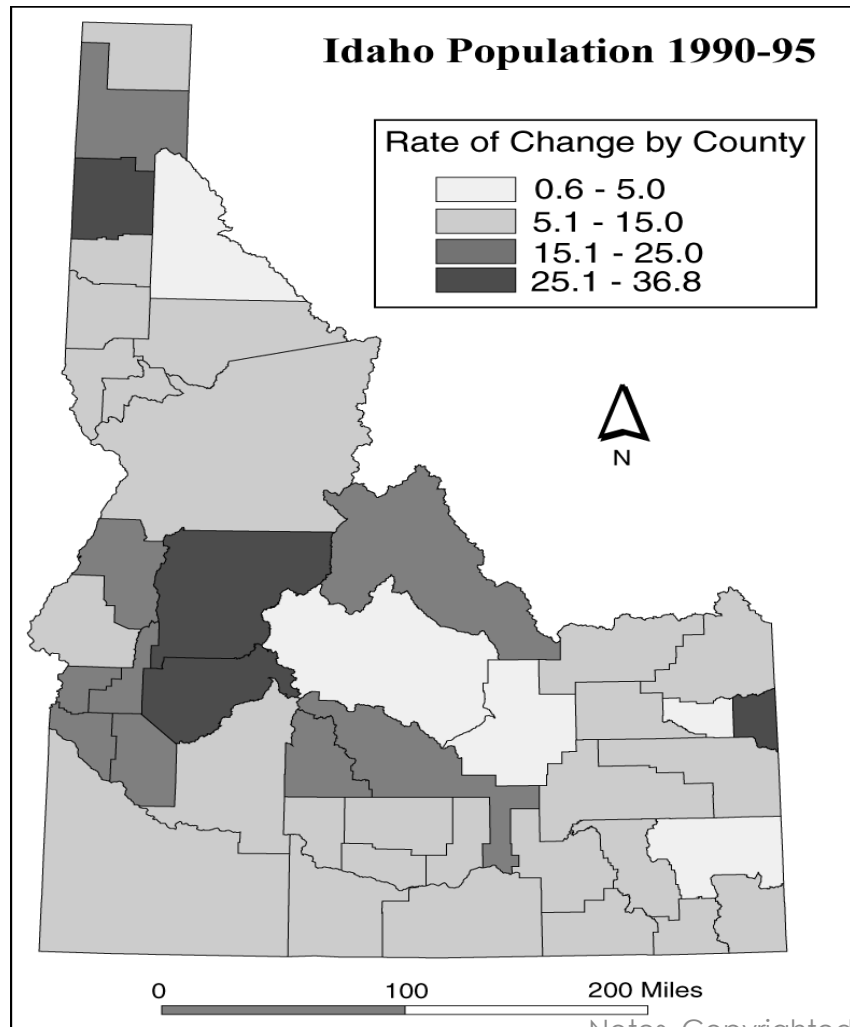
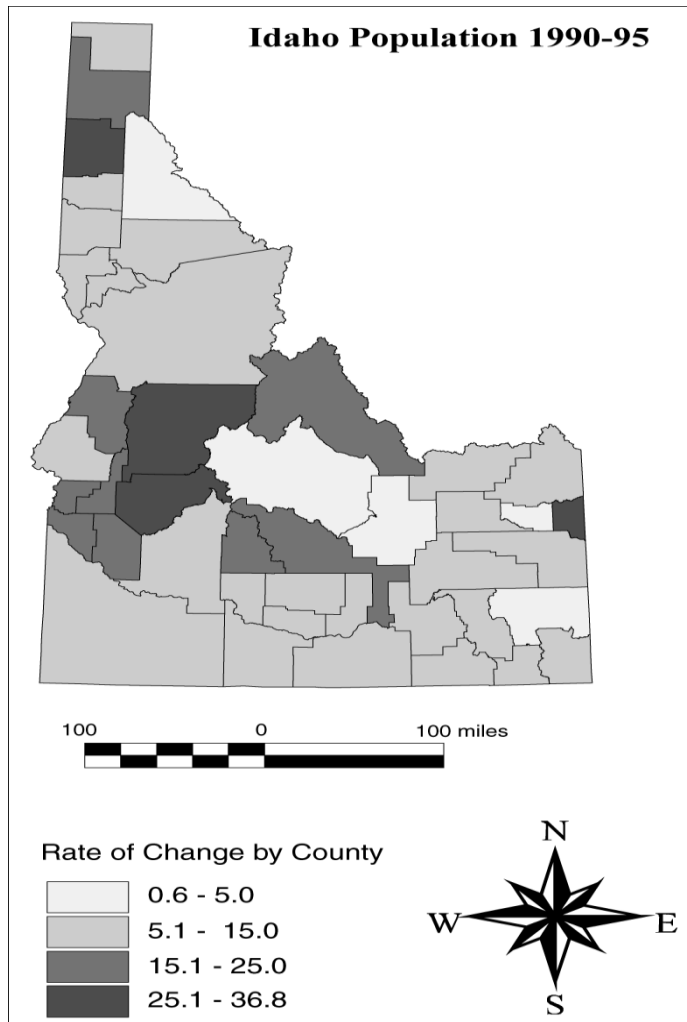


Figure 10.15

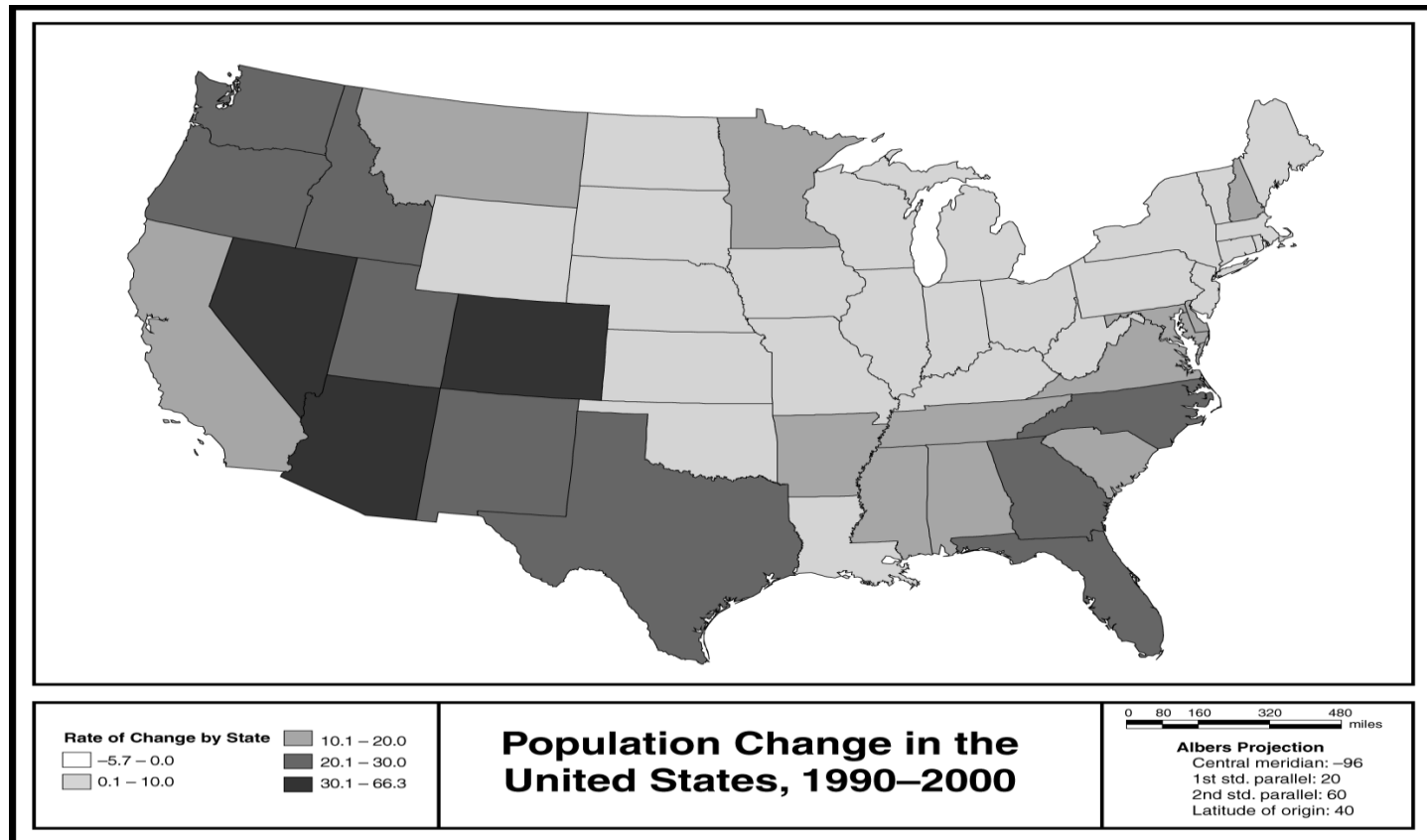
Use a box around the legend to draw the map reader's attention to it.

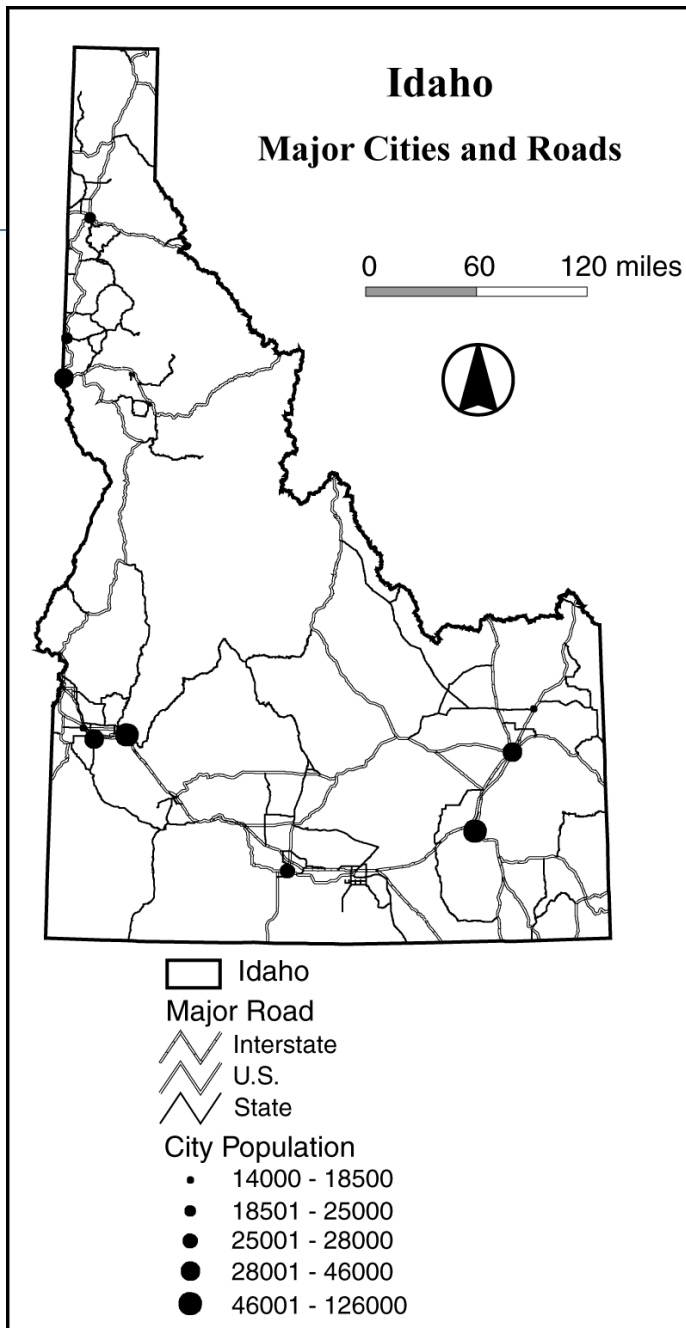
Too many details at the bottom



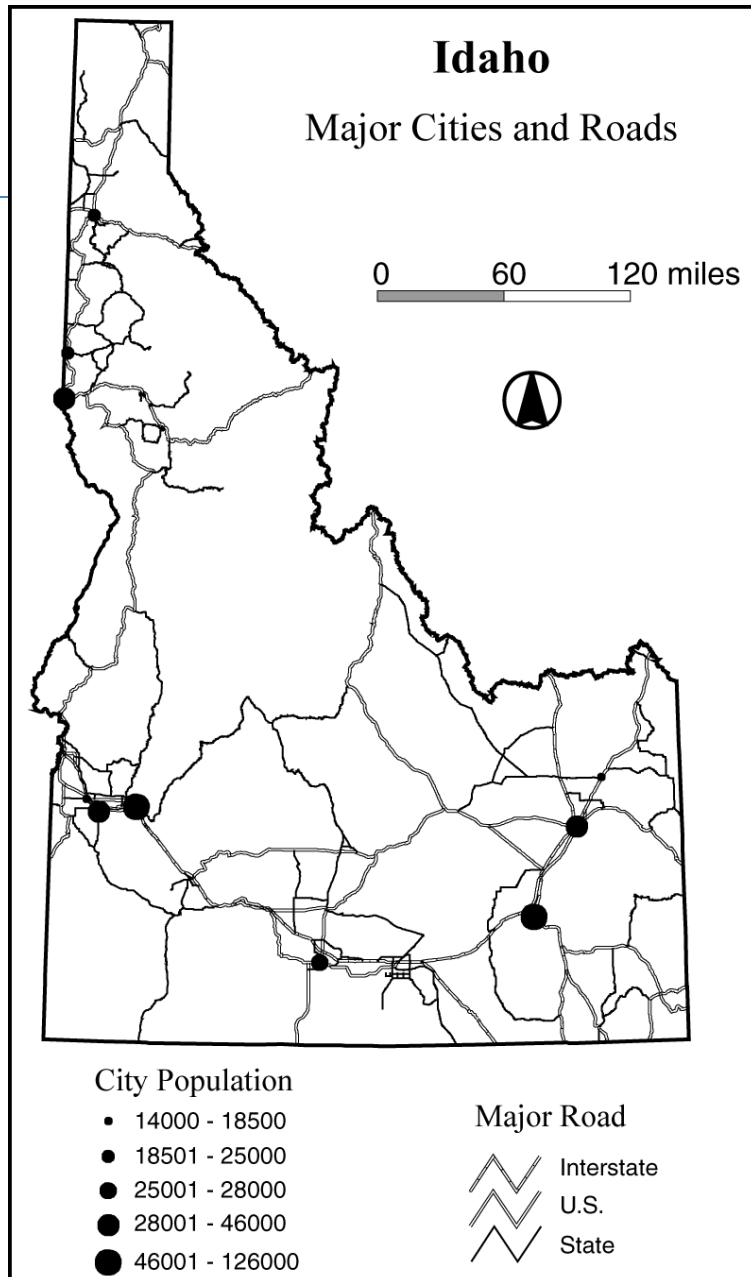
A poorly balanced map.

Layout template in ArcMap



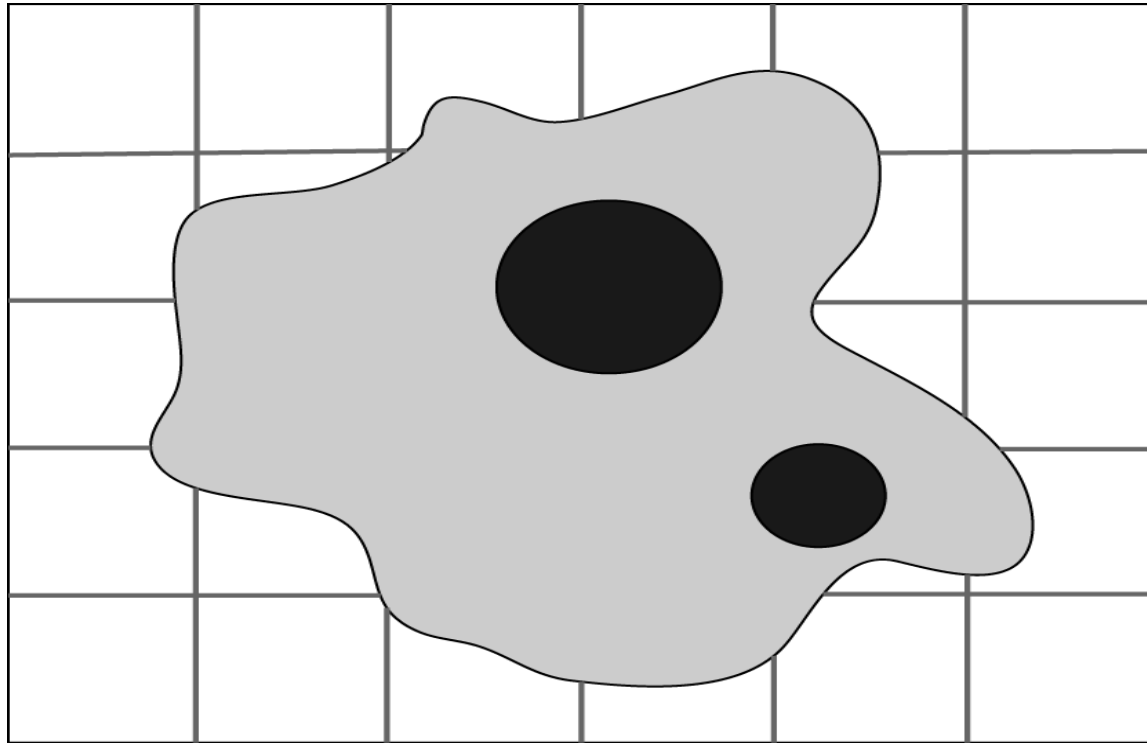


A lengthy legend is confusing and can create a problem in layout design.



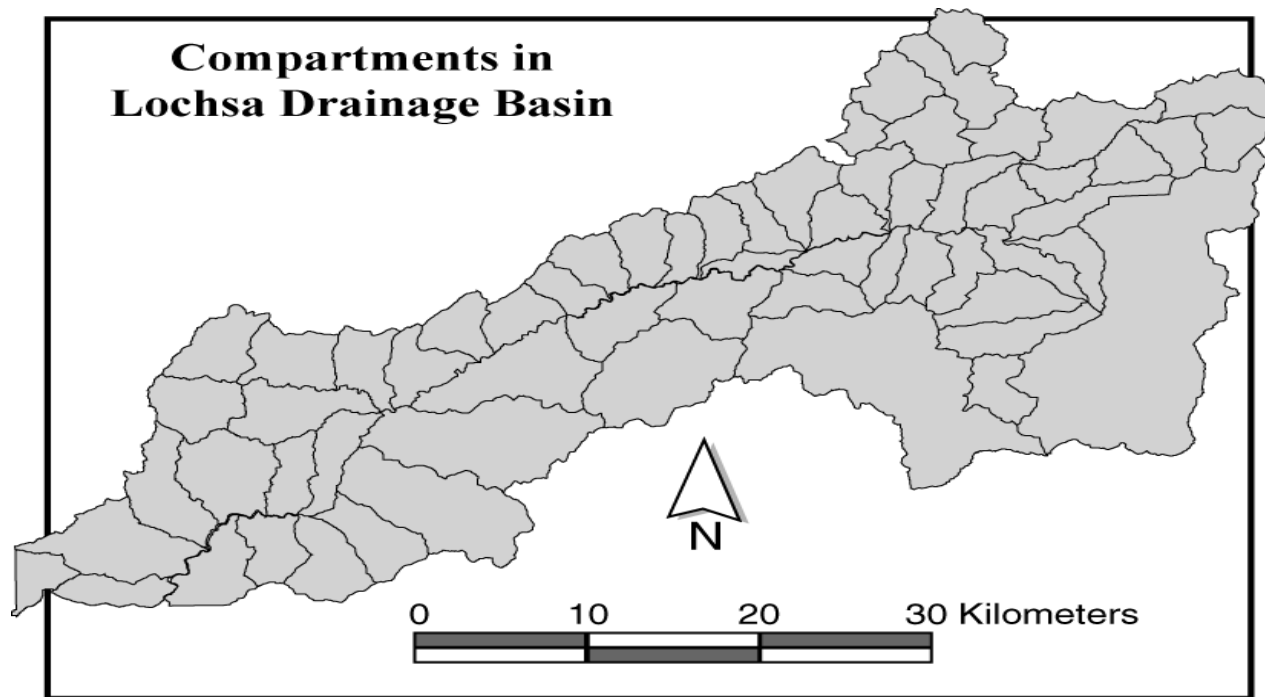
The lengthy legend is now separated into two parts. Also, the unnecessary outline symbol is removed from the legend.

Visual Hierarchy

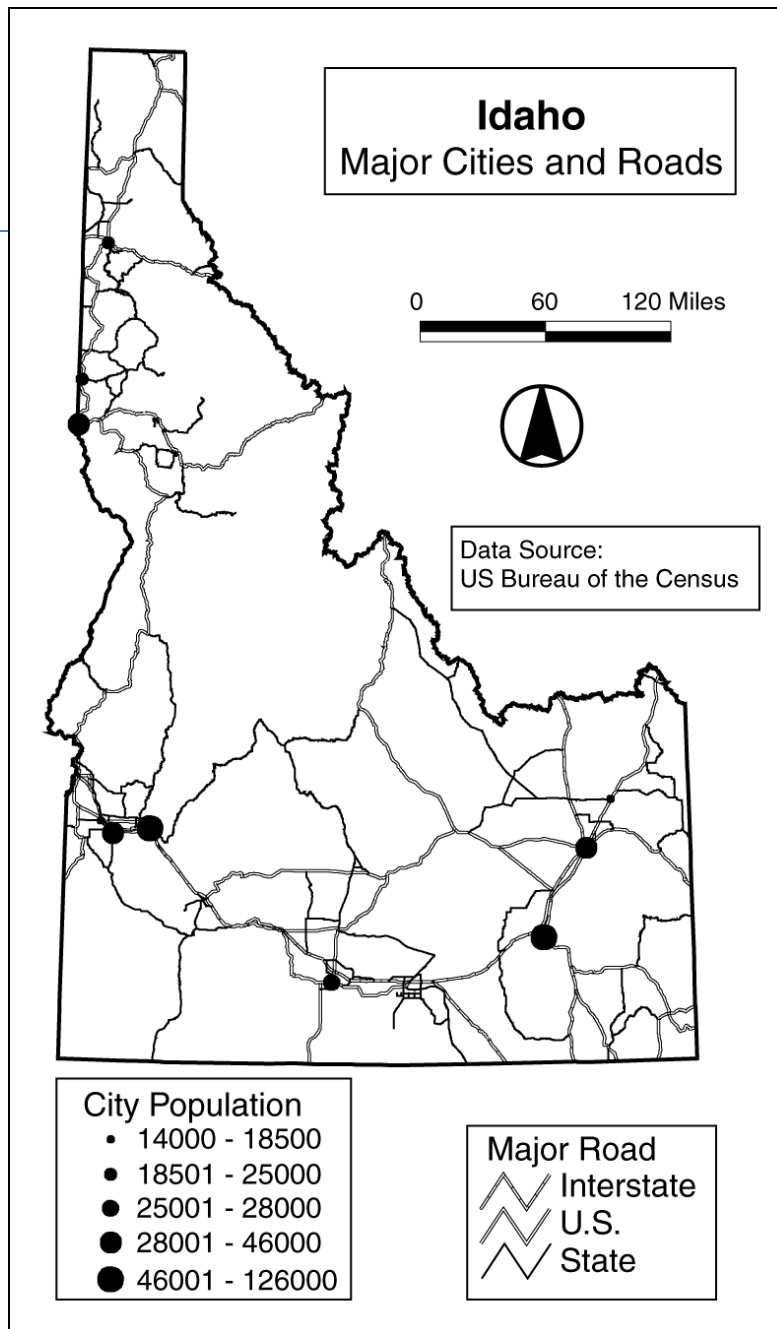


Visual hierarchy example. The two black circles are on top (closest to the map reader), followed by the gray polygon and the grid.

Interposition



The interposition effect in map design.

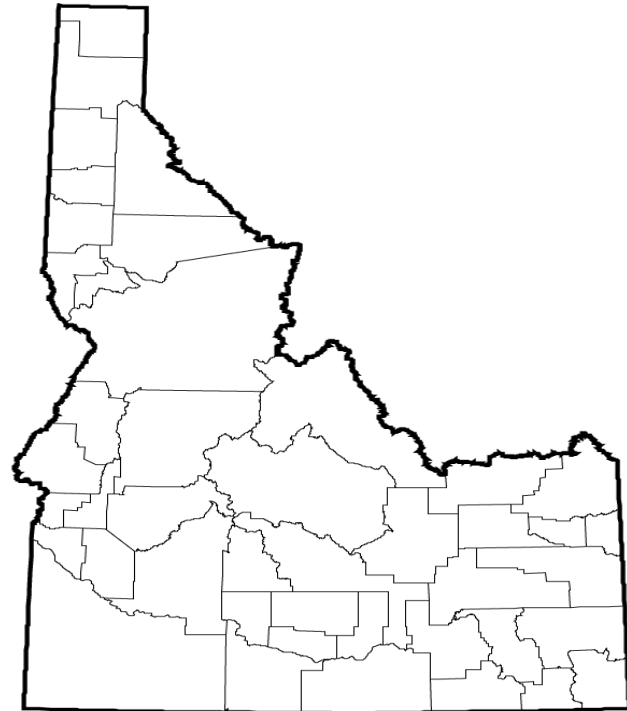


A map looks confusing if it uses too many boxes to highlight individual elements.

Contrast Characteristics



(a)



(b)

Contrast is missing in (a), whereas the line contrast makes the state outline look more important than the county boundaries in (b).

Spatial-Analysis Assisted Hydrologic and Environmental Modeling and Water Resources Management

- Topics
 - Spatial inputs to hydrologic and environmental modeling efforts. Inputs to lumped and distributed hydrologic models.
 - Generation of gridded precipitation data, interpolation (examples of radar-based precipitation datasets),
 - Derivation of watershed-specific properties for hydrologic and environmental modeling (pollutant load),
 - Watershed delineation using D-8 algorithm,
 - Processing of land use and land cover data sets.
 - Several examples of spatial analysis-based hydrologic and environmental modeling.
 - Linking of geographic information systems and water management models.

Watershed Concept

- From a civil engineering perspective, the most important subsystem is the **watershed** system. The watershed system may be conceptualized as a spatial subset of the lithospheric system. The watershed is simply that area of land that contributes surface runoff to a common point of interest

Topographic Maps

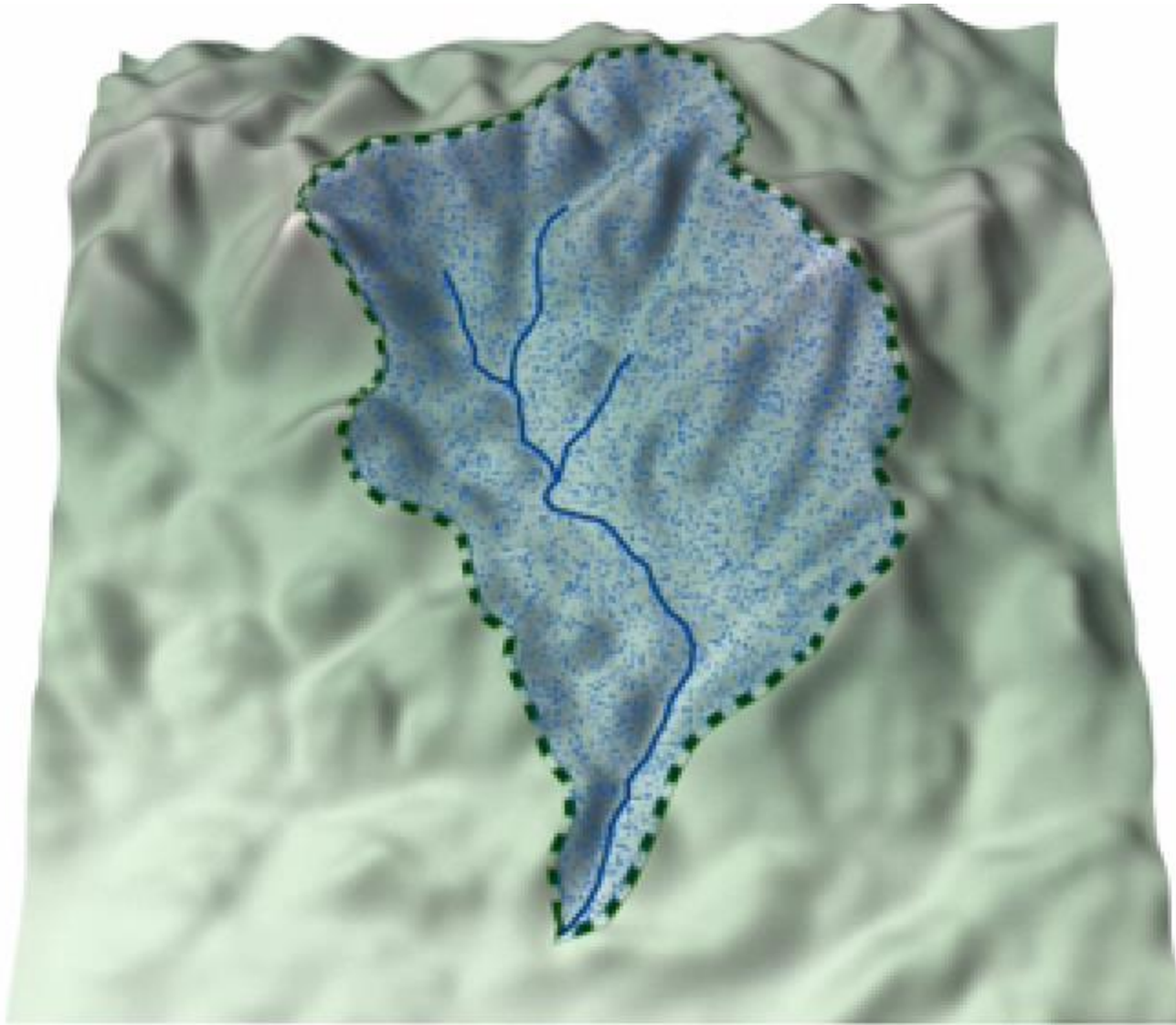
- Reading Topographic Maps
 - Each contour line on a topographic map represents a ground elevation or vertical distance above a reference point such as sea level.
 - All points along any one contour line are at the same elevation
 - If a hill is more or less circular then the map will show it as a series of more or less concentric circles (contour lines)

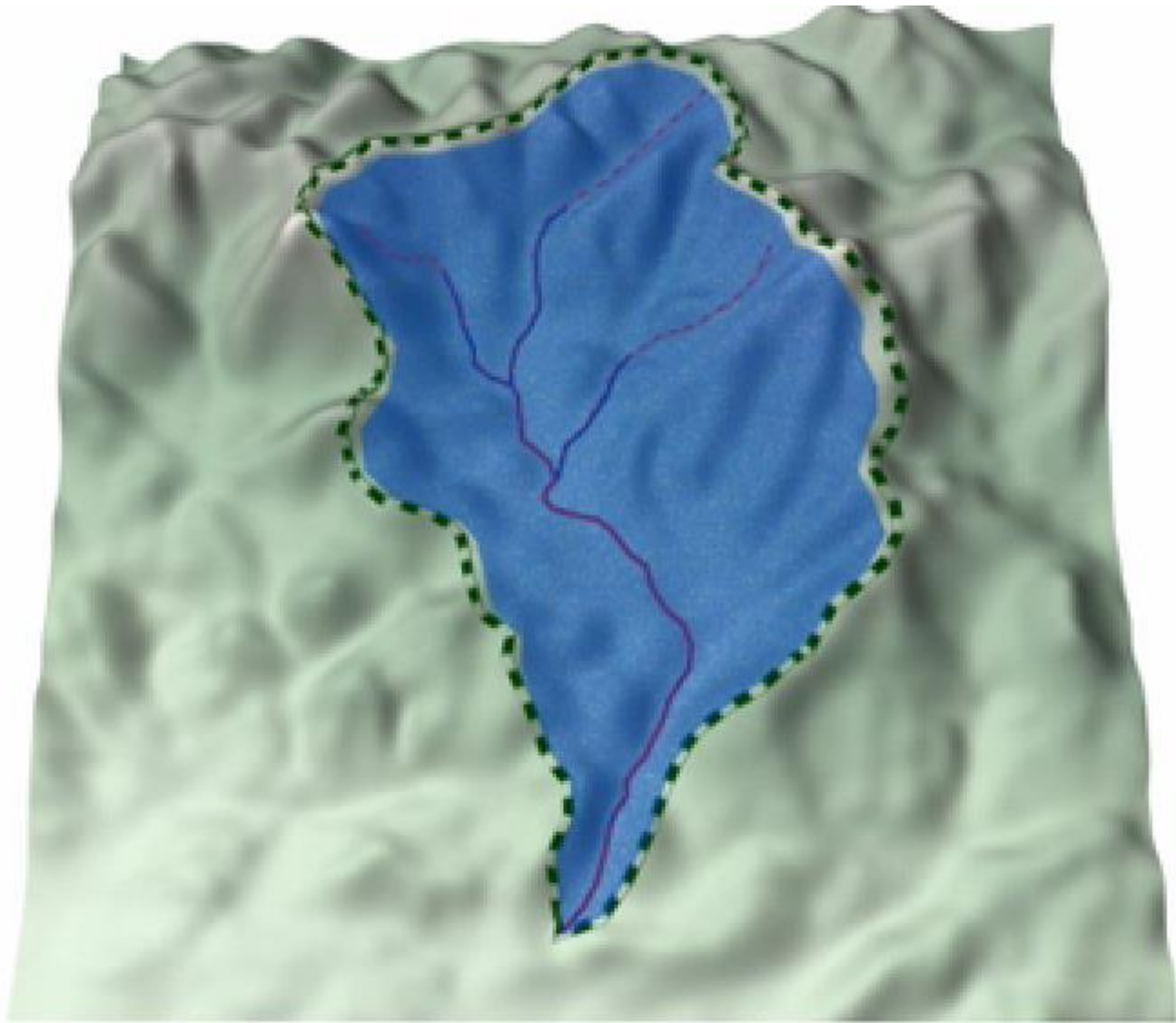
Watershed Delineation

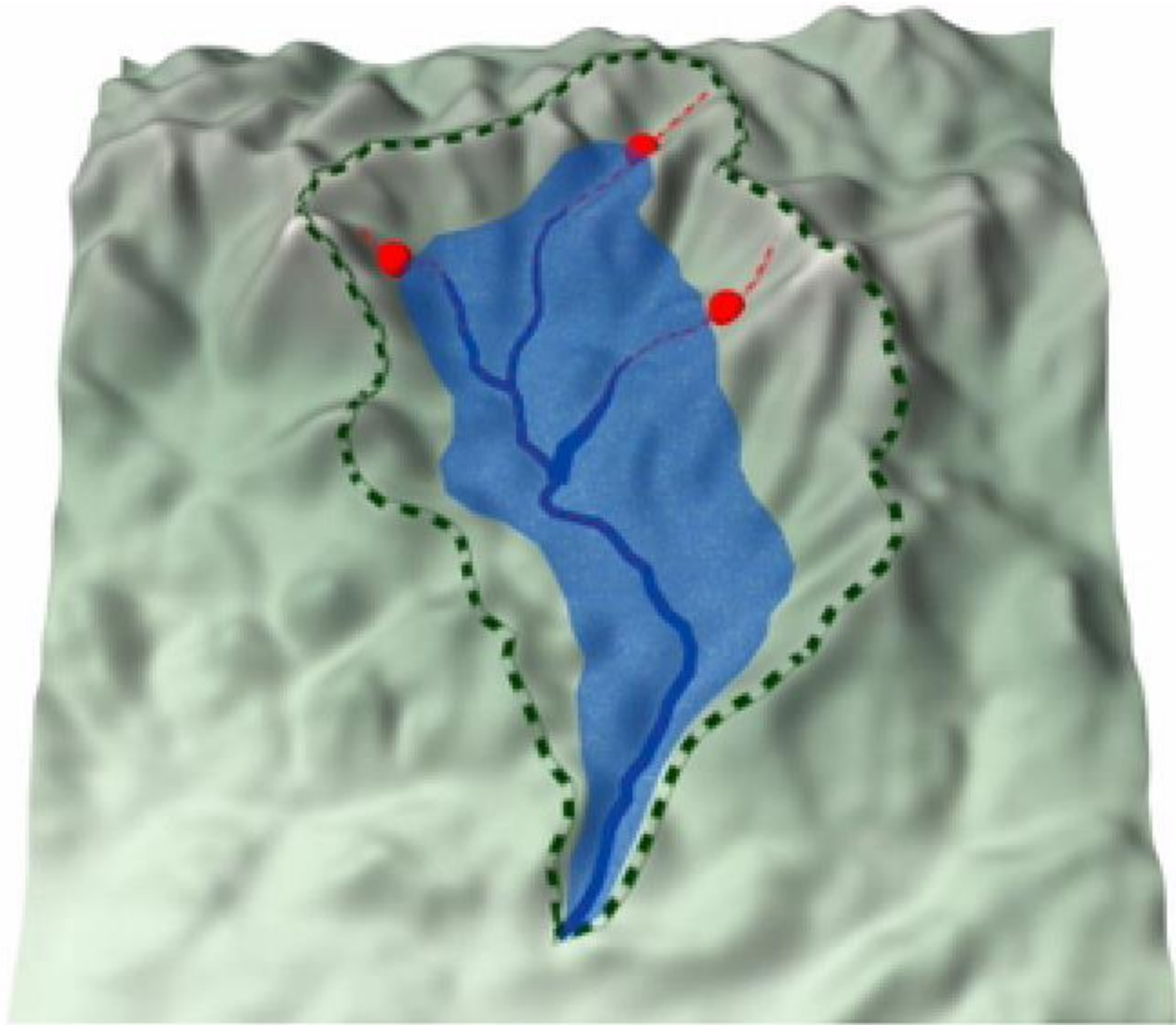
- Identify point of interest on a stream or channel.
- Obtain topographic map containing point.
- Start at point of interest and begin drawing a line in the upstream direction of stream or channel always bisecting contours at a right angle.
- Continue drawing line until you return to the original point.

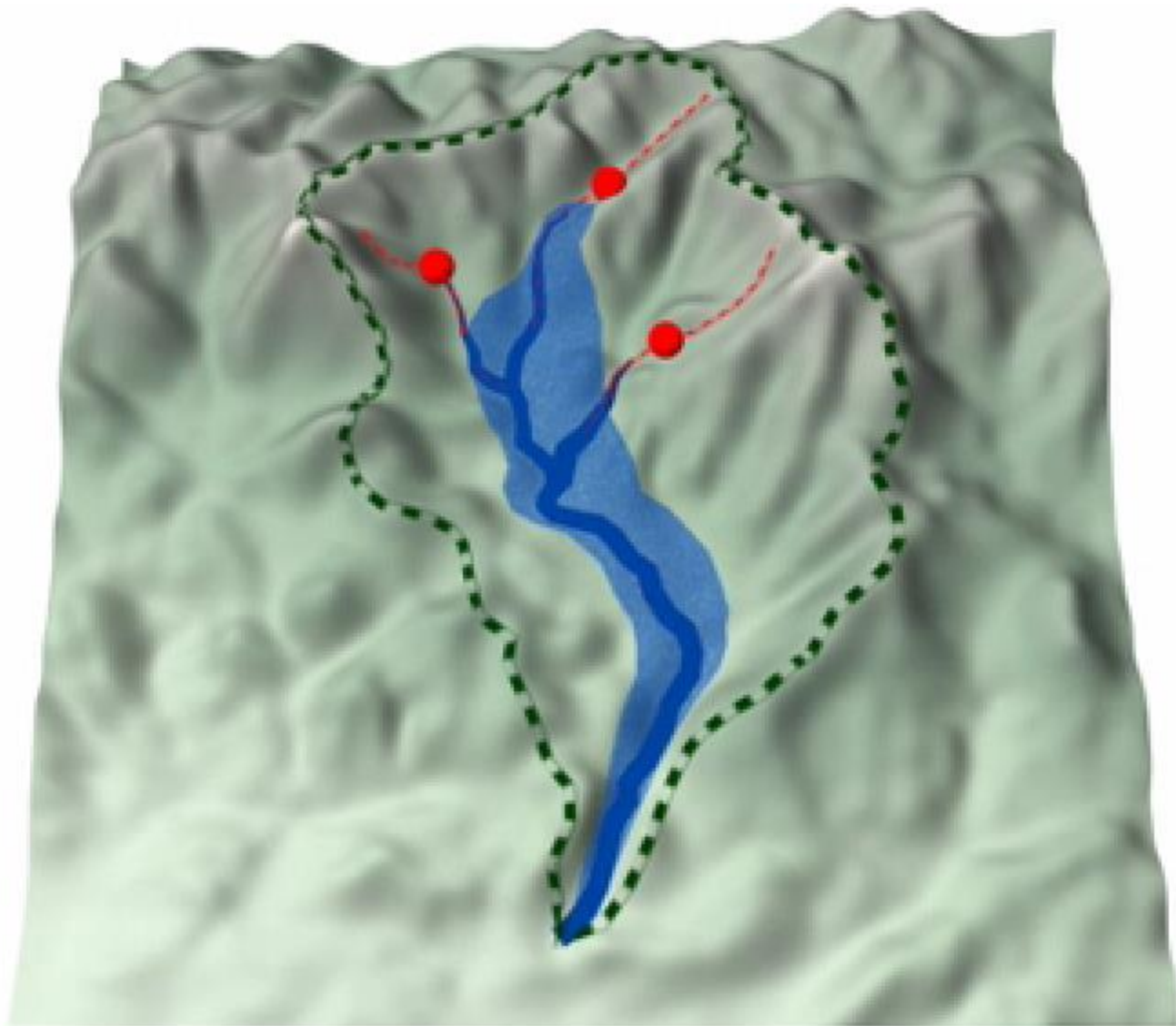
Simple rules for watershed delineation

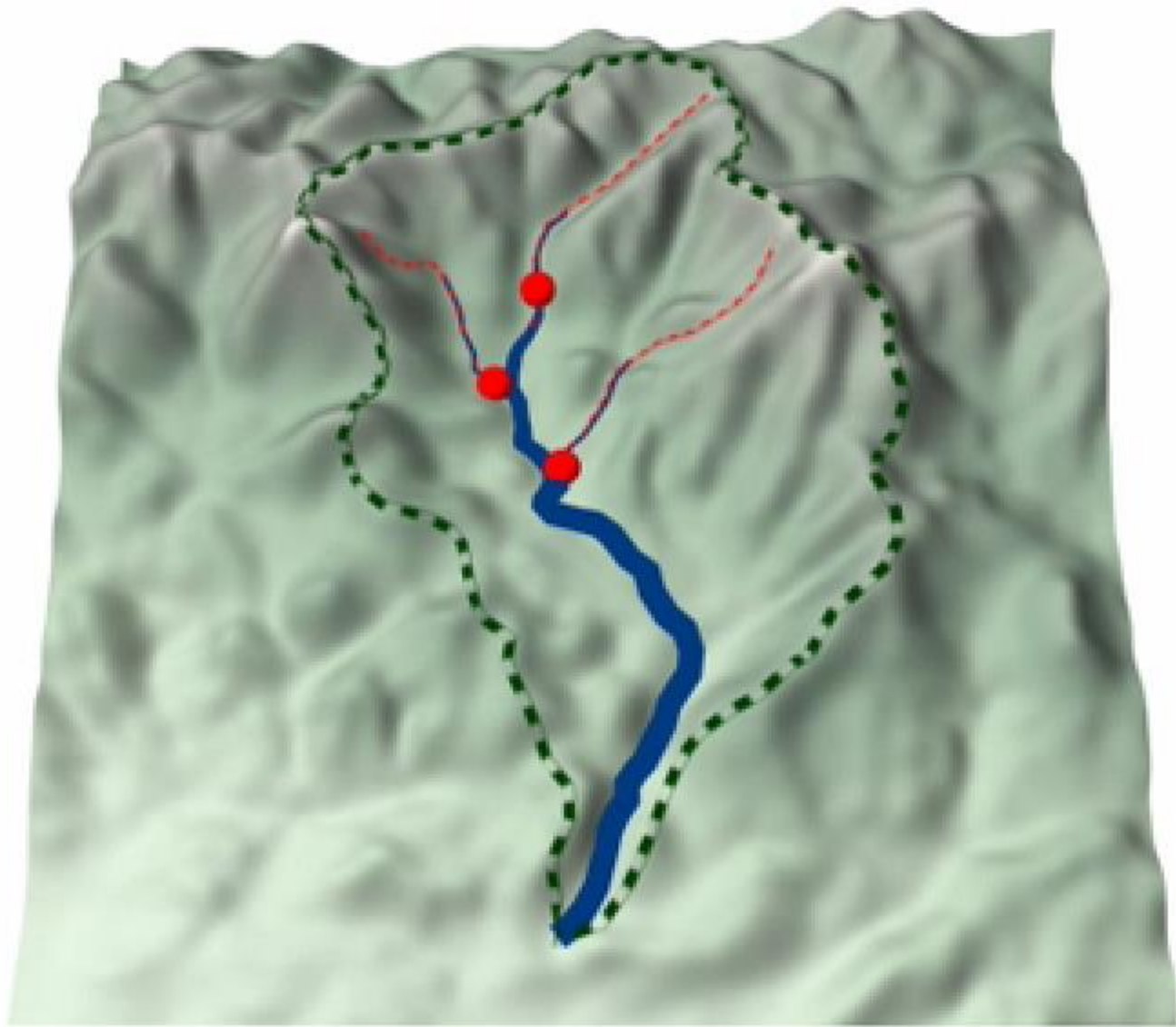
- Surface water generally flows perpendicular across contour lines
- Ridges are indicated by the highest elevation contour line in the area
- Drainages are indicated by contour V's pointing upstream

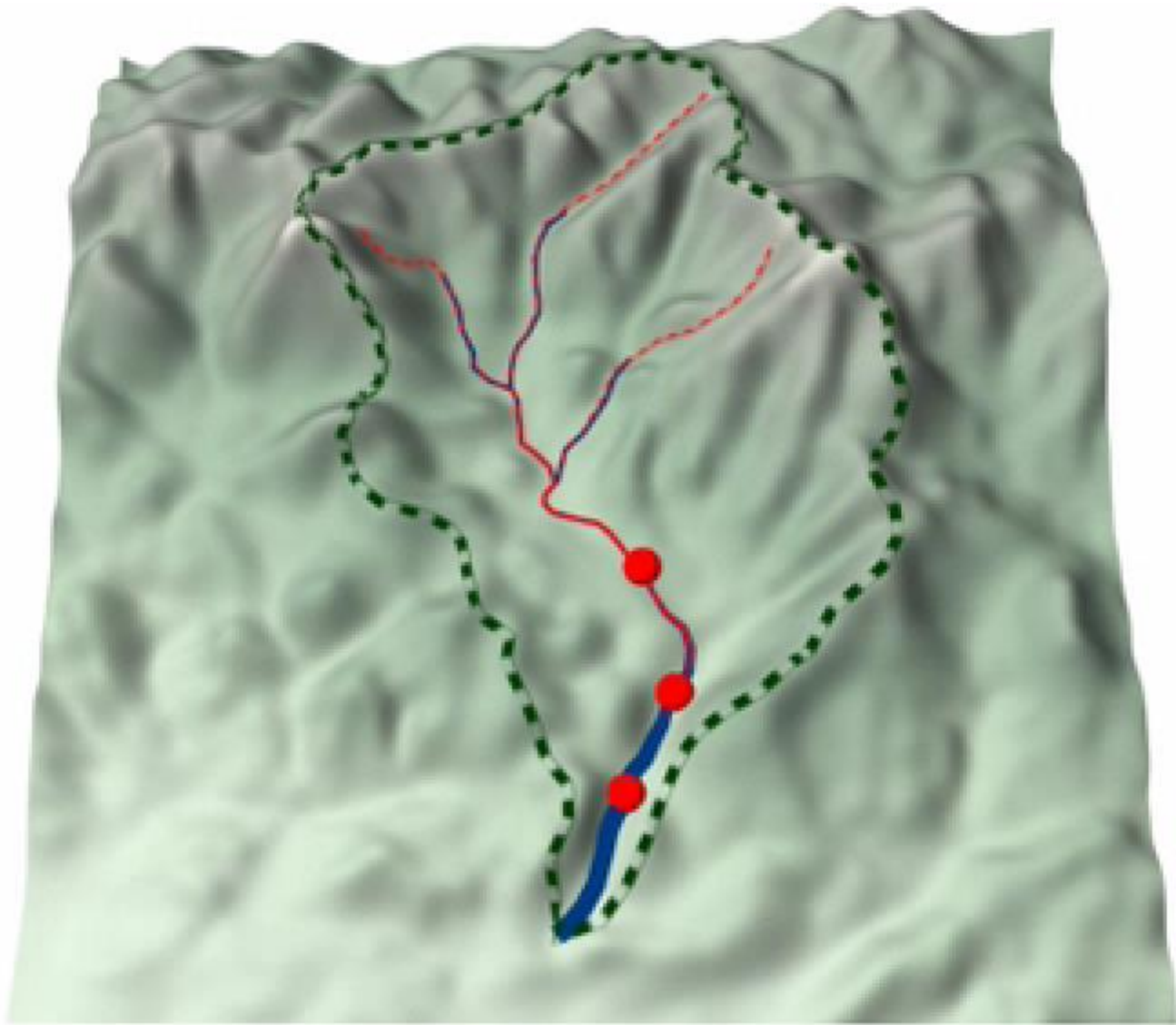


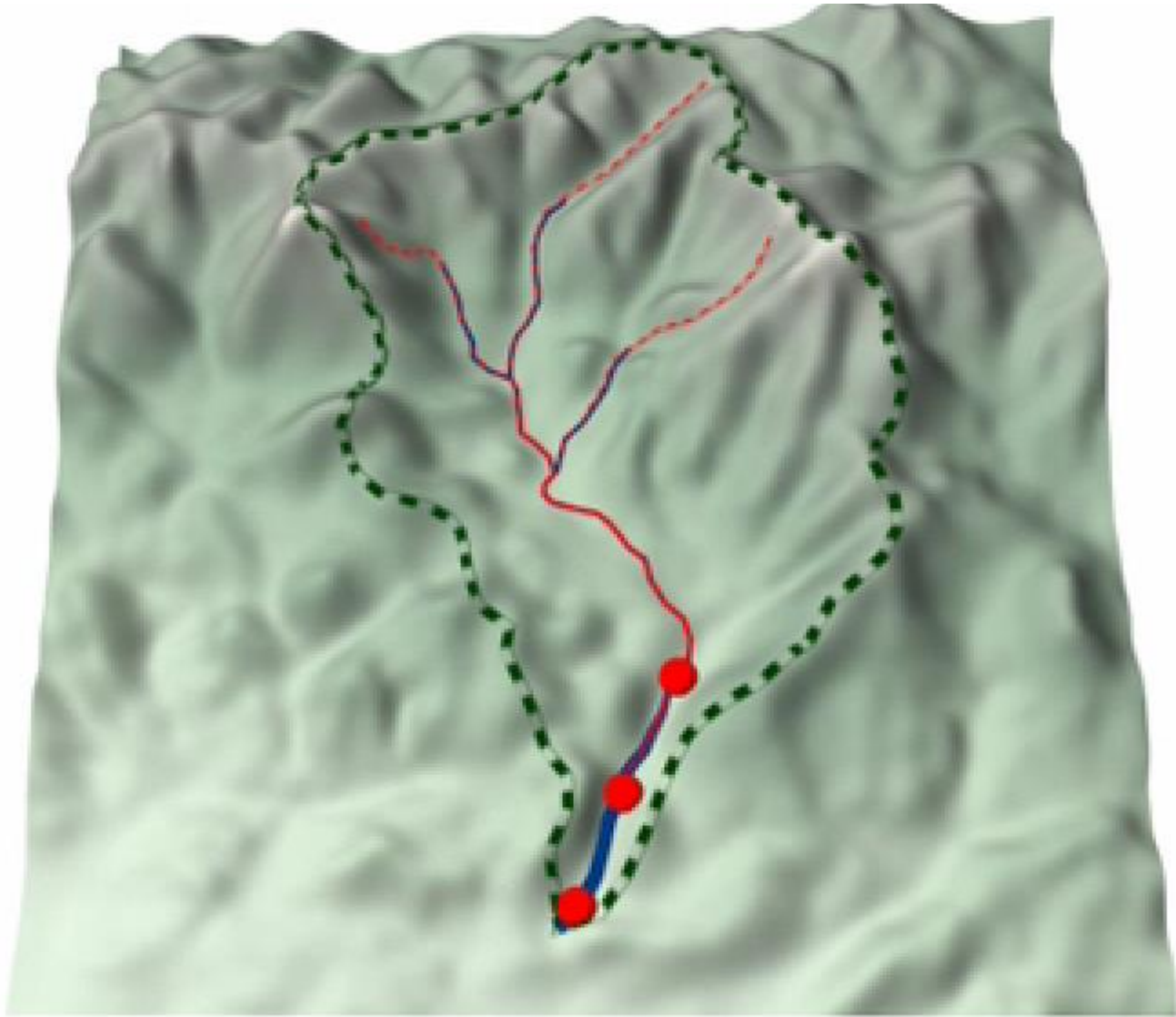


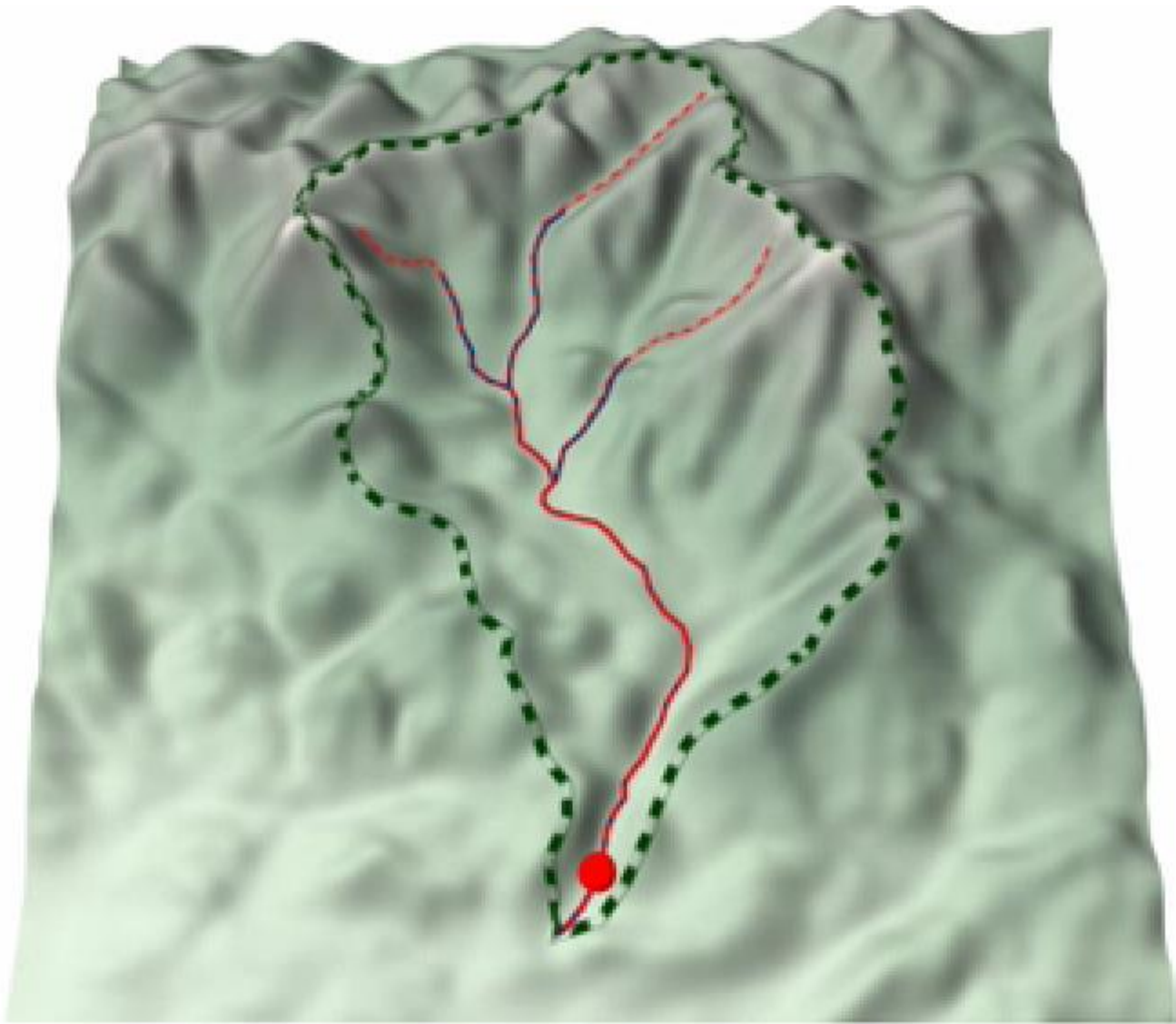






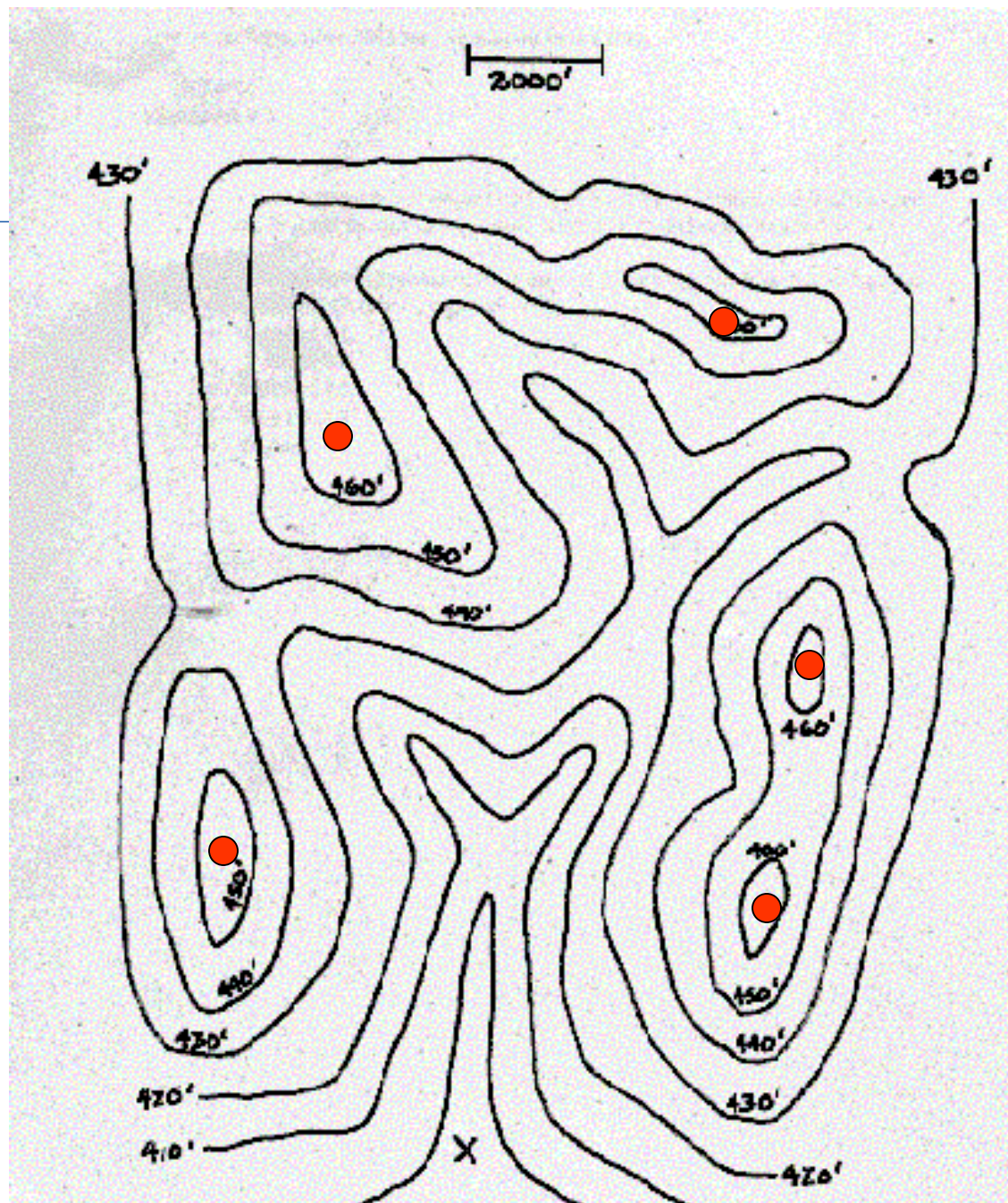






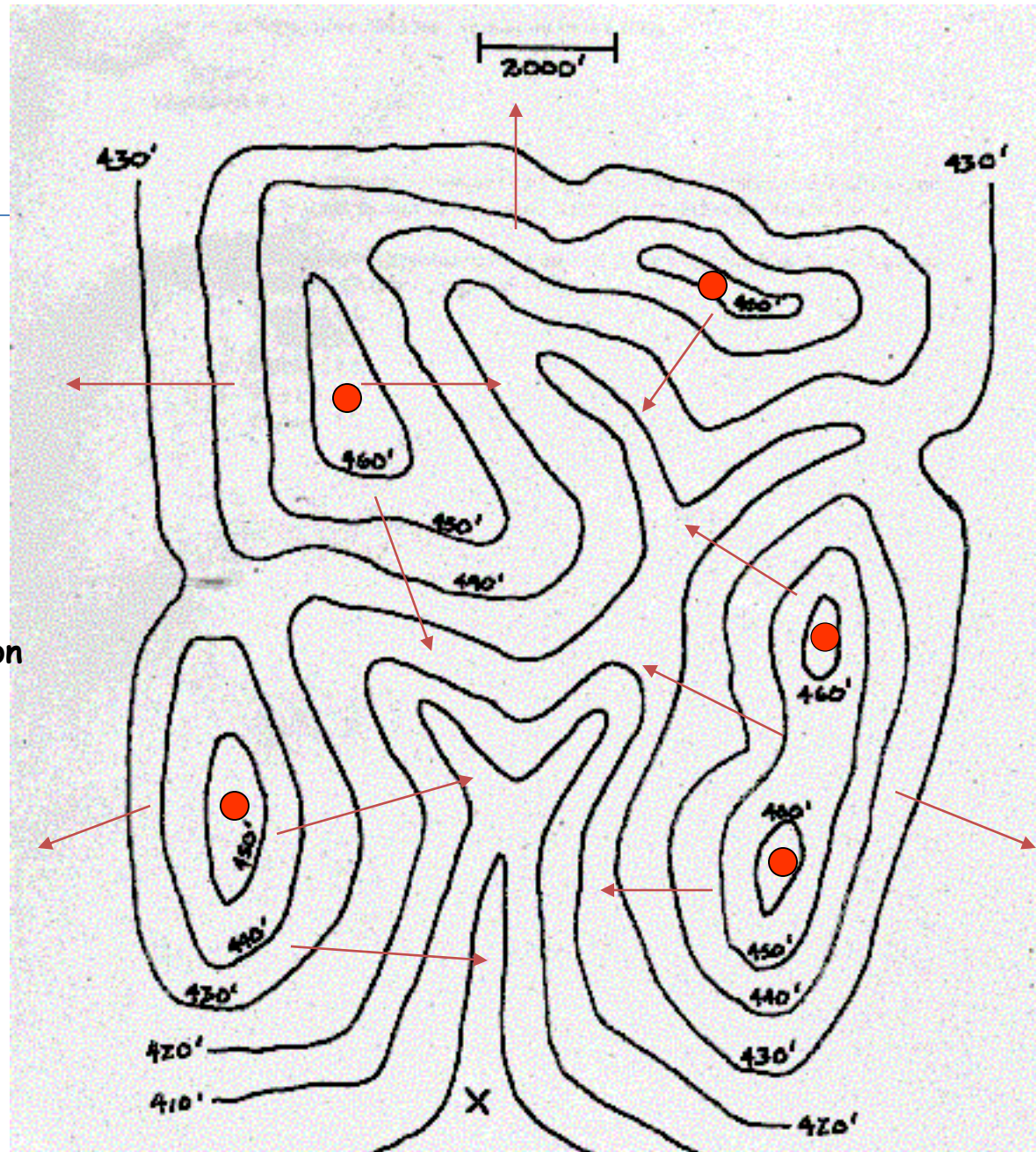
Watershed Delineation

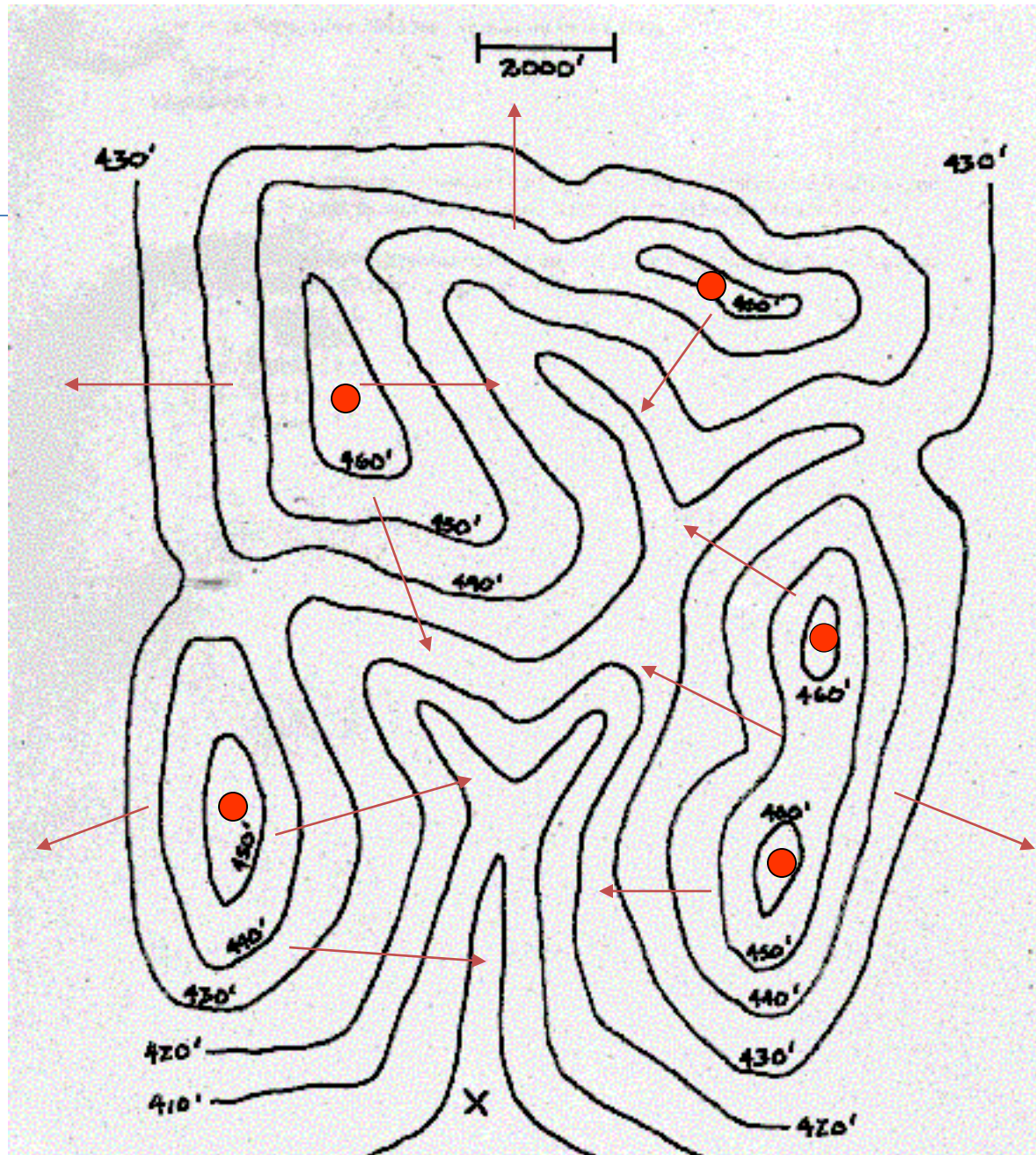
Example

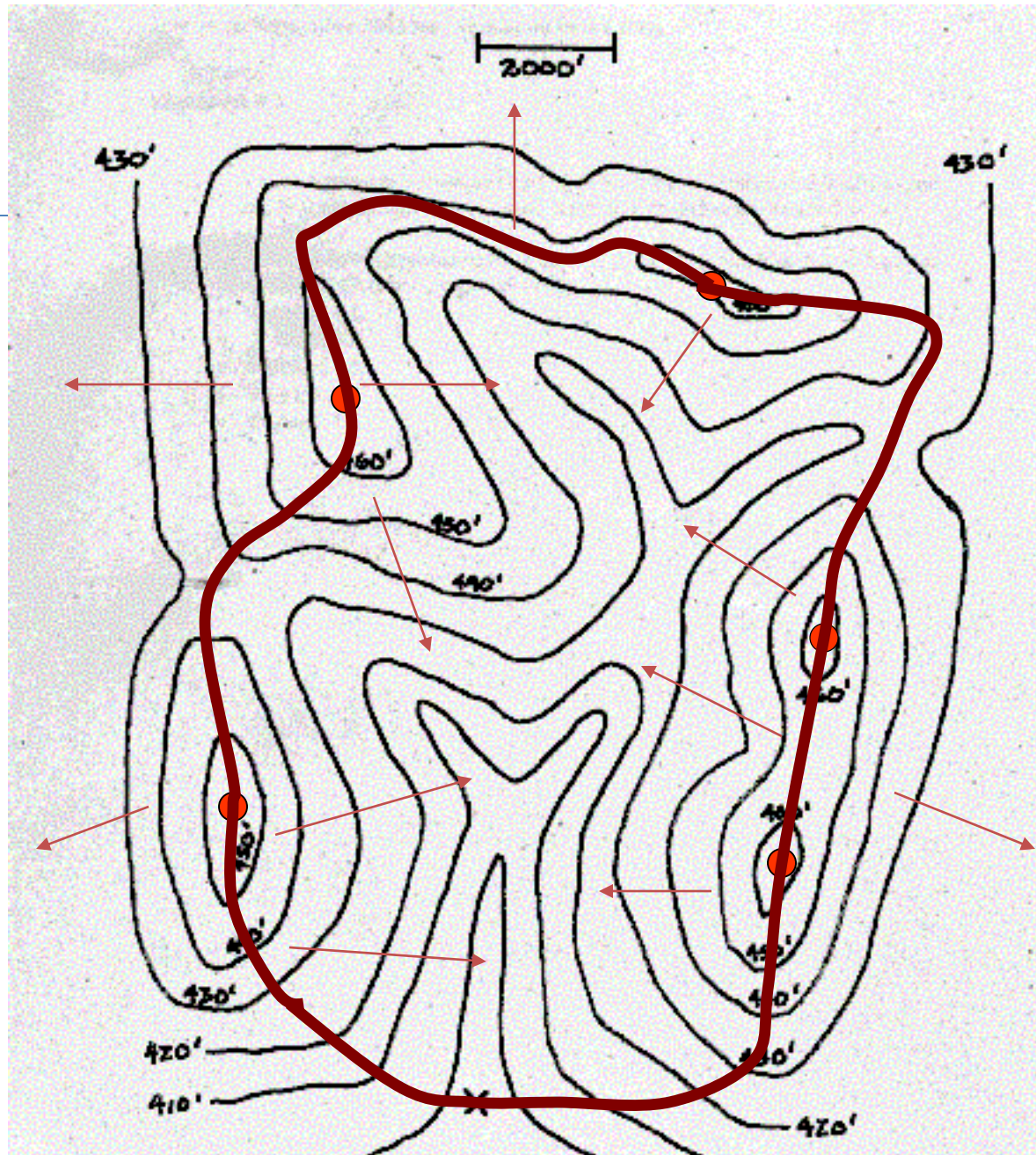


Identify
the highest
points and
associated
contours

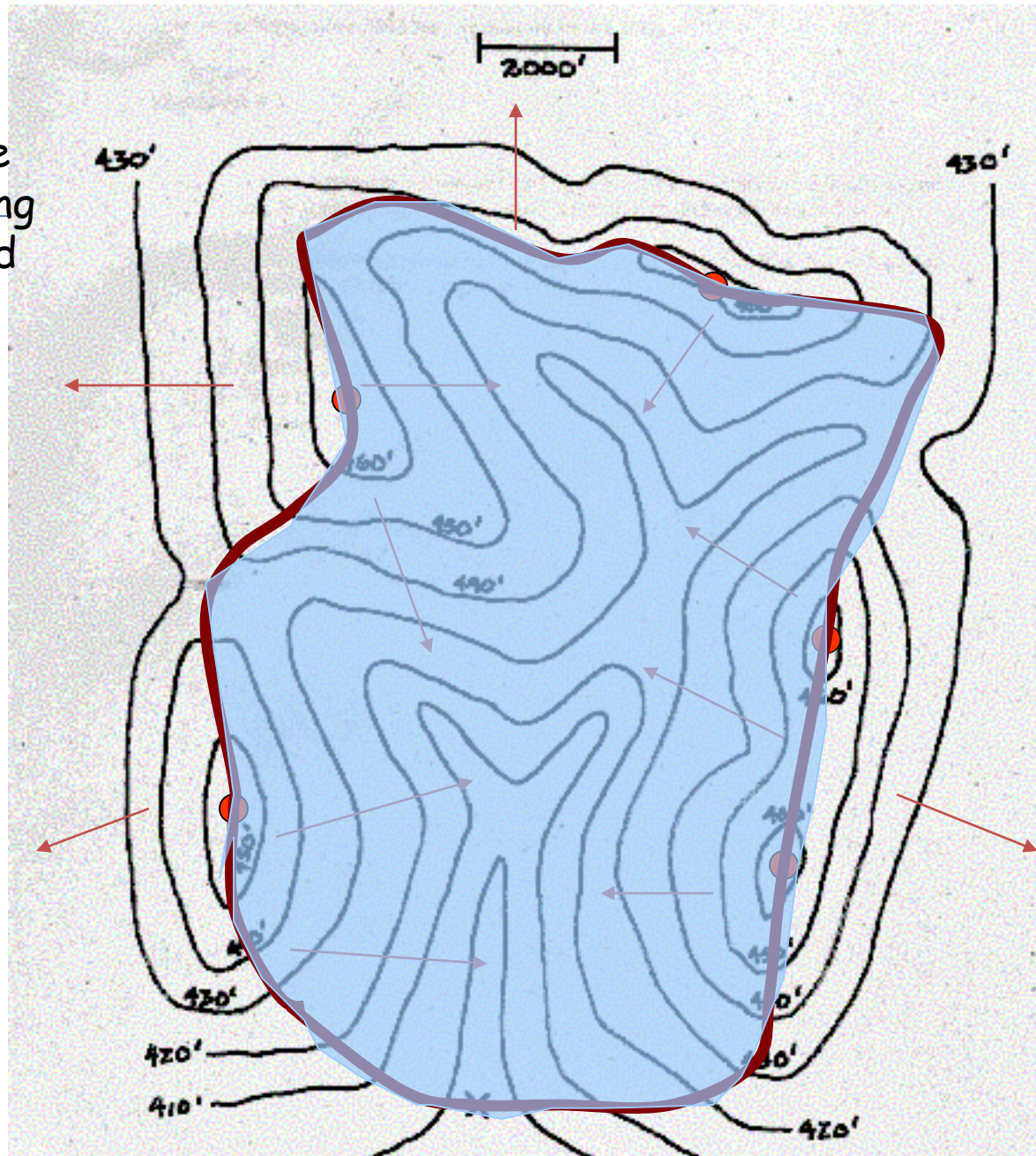
Understand the movement of water if a drop of water was to be dropped at a point and it's travel direction



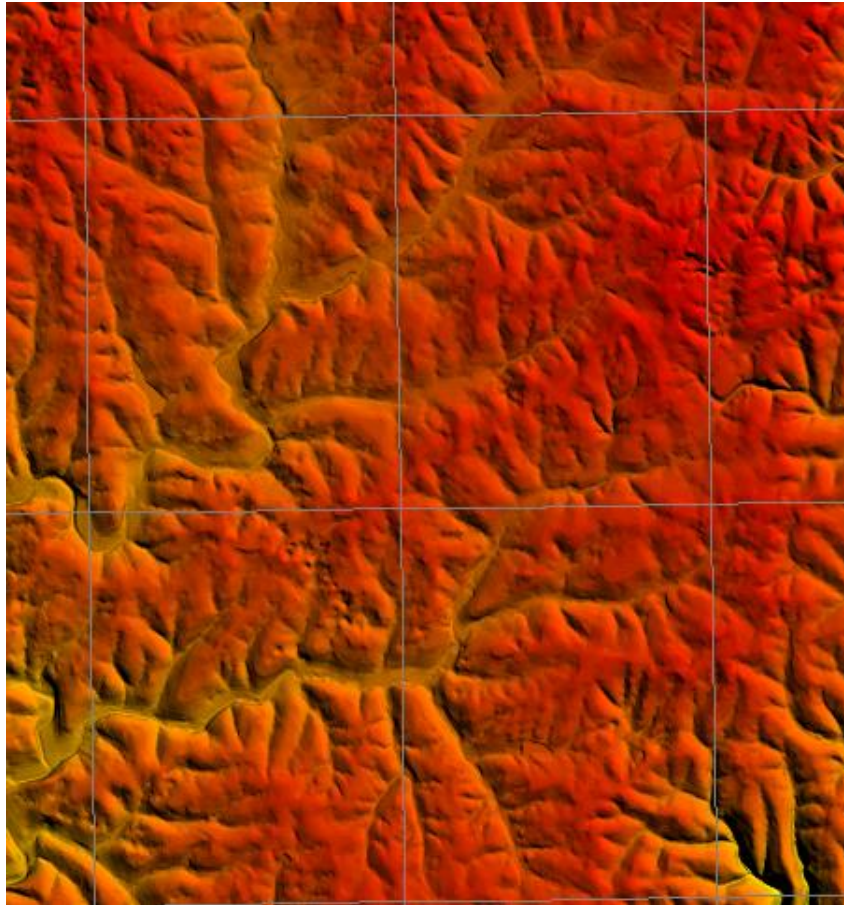




Approximate
the area using
the scale and
grid

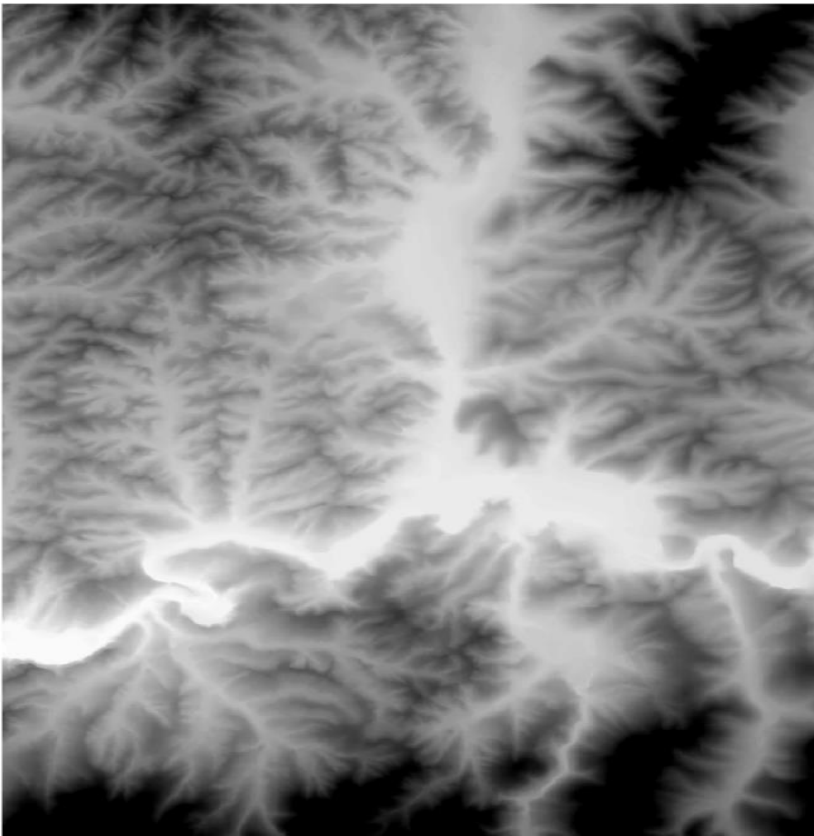


Digital Elevation Model



Raster Map

Raster maps are cell-based. They can be qualitative (categorical) or quantitative (numeric).



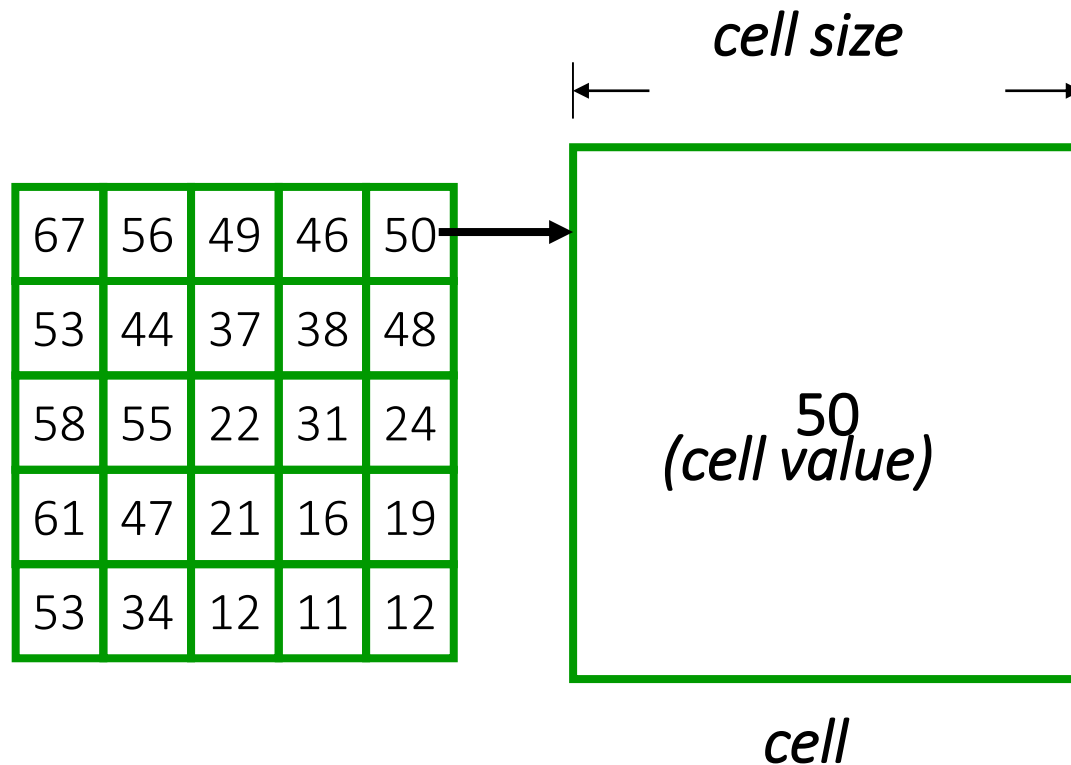
Map showing raster-based elevation data. Cells with higher elevations have darker shades.

DEM

- A DEM is a raster representation of a continuous surface, usually referencing the surface of the earth.
- The accuracy of this data is determined primarily by the resolution (the distance between sample points).
- Other factors affecting accuracy are data type (integer or floating point) and the actual sampling of the surface when creating the original DEM

Cell Definition

- Cell Size



DEM errors

- Errors in DEMs are usually classified as either sinks or peaks. A sink is an area surrounded by higher elevation values and is also referred to as a depression or pit.
- This is an area of internal drainage. Some of these may be natural, particularly in glacial or karst areas (Mark, 1988), although many sinks are imperfections in the DEM.
- Likewise, a spike or peak is an area surrounded by cells of lower value. These are more commonly natural features and are less detrimental to the calculation of flow direction.

DEM imperfections

- Errors such as these, especially sinks, should be removed before attempting to derive any surface information. Sinks, being areas of internal drainage, prevent downslope flow routing of water.
- The number of sinks in a given DEM is normally higher for coarser resolution DEMs. Another common cause of sinks results from storing the elevation data as an integer number. This can be particularly troublesome in areas of low vertical relief.
- It is not uncommon to find 1 percent of the cells in a 30-meter-resolution DEM to be sinks.

Watershed Delineation using DEM

- Digital elevation models (DEMs) along with spatial analysis tools are commonly used for automatic delineation of watersheds as they are **superior to manual delineation** of watershed boundaries.
- The automatic derivation of topographic watershed data from DEMs is faster, **less subjective**, and provides more **reproducible measurements** than traditional manual techniques applied to topographic maps

Topographic Maps

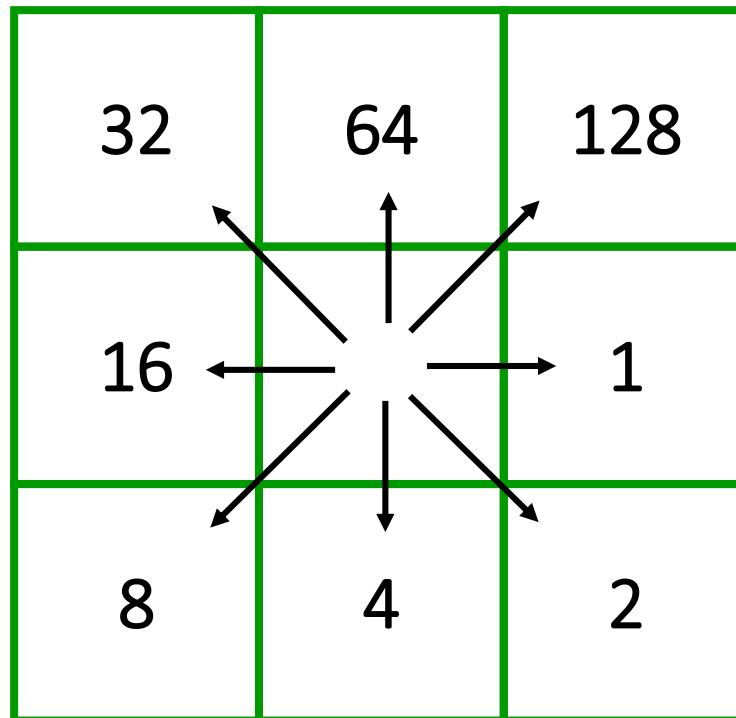
- **Topographic maps** are the traditional way of representing land surface terrain and streams
- **Watersheds** can be hand-delineated from these maps
- **DEM's** of equivalent accuracy are now available for most map series in the US

Automatic Delineation

- Several public domain and proprietary spatial analysis software provide tools for Watershed Delineation
- ArcGIS
- BASINS (Better Assessment Science Integrating point and Non-point Sources)
 - Requires ArcView
- BASINS 4 (ArcView or ArcMap independent)
- TAUDDEM (Utah State University)

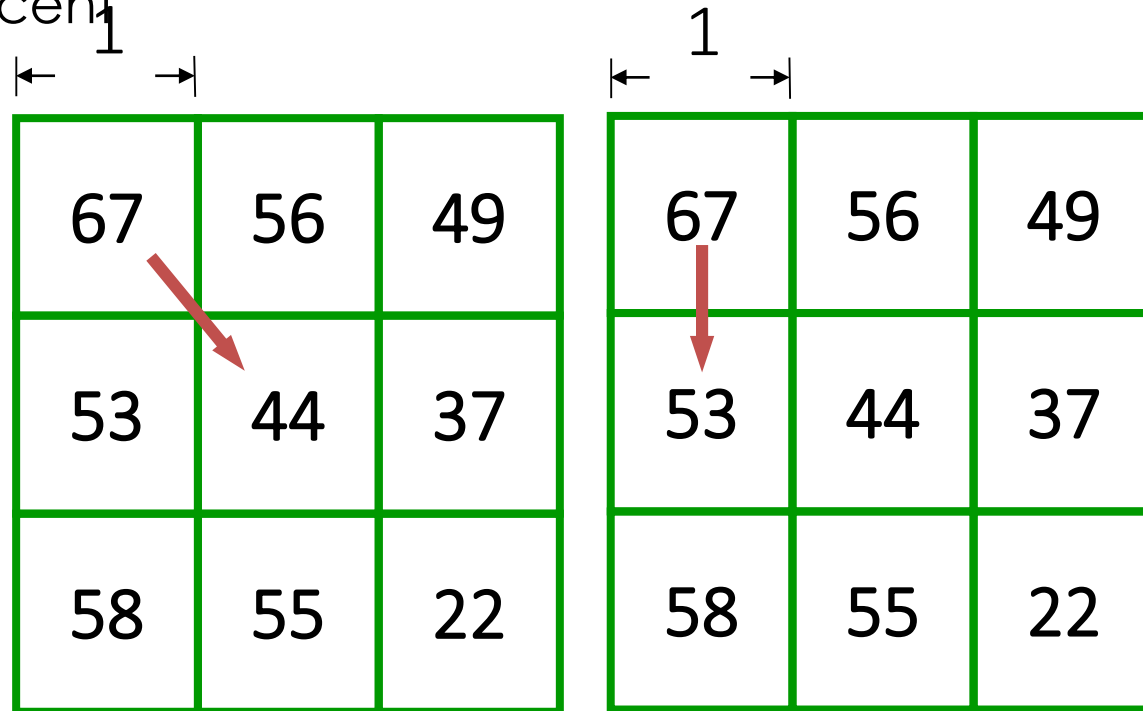
8-Point Pour Algorithm (D-8)

- Direction of Flow



Direction of Steepest Descent

- Steepest Descent

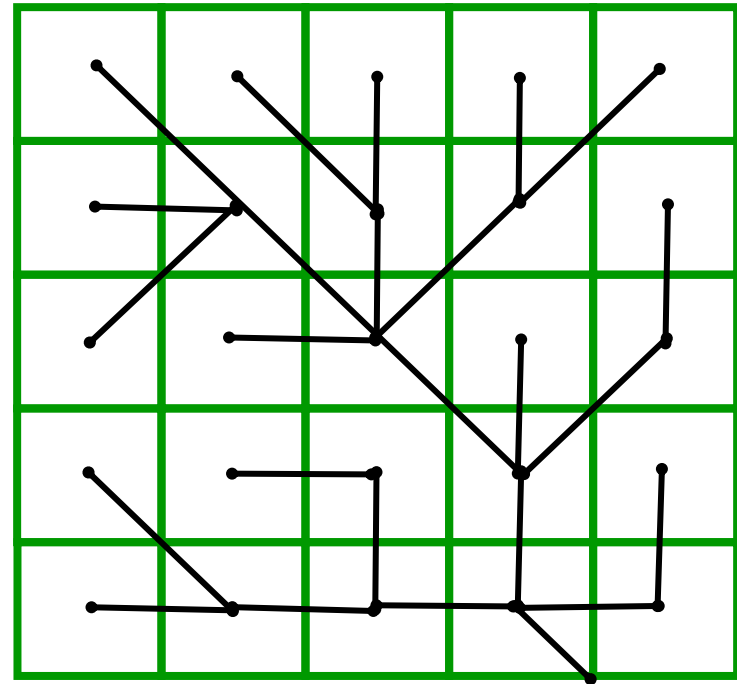
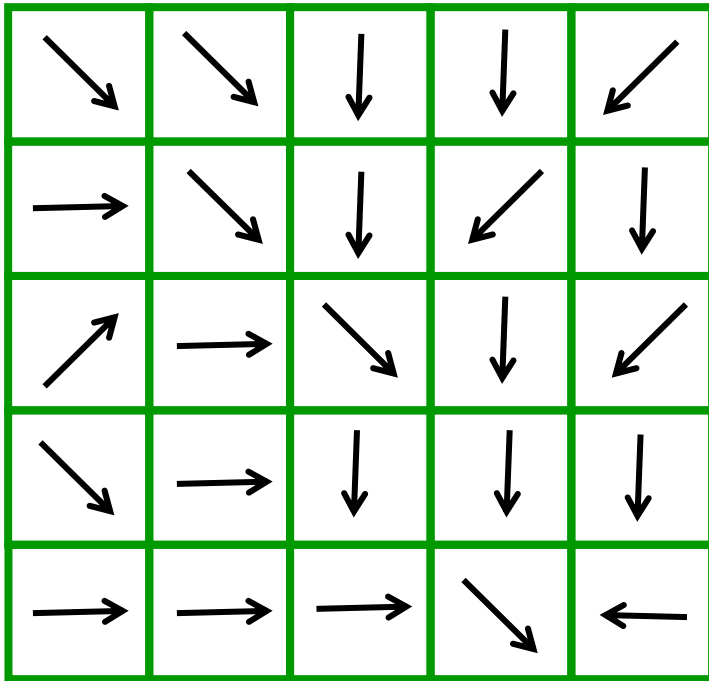


Slope: $\frac{67 - 44}{\sqrt{2}} = 16.26$

$$\frac{67 - 53}{1} = 14$$

Grid Network

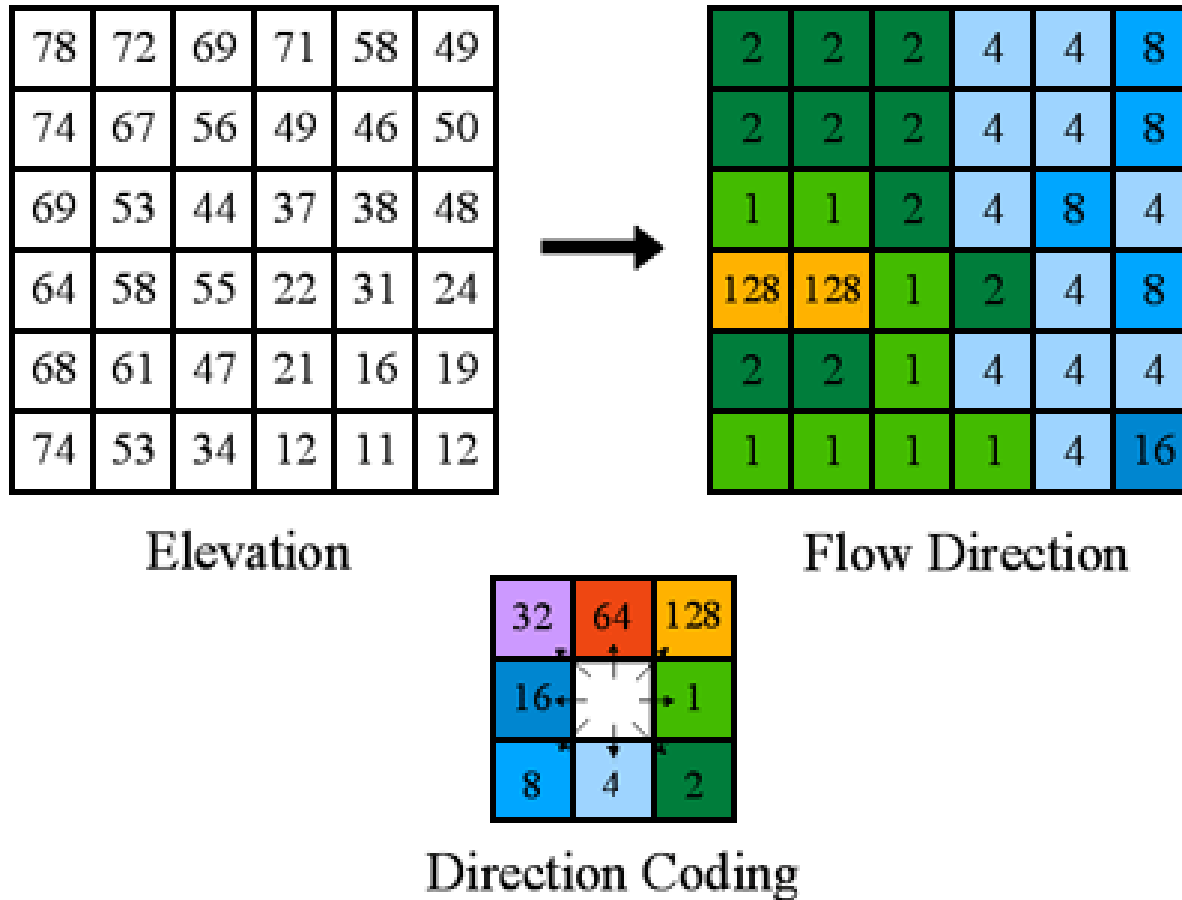
- Flow



Eight Direction Pour Point Model

- DEM cell elevation is at the **cell center**
- **Eight direction pour point** model leads to flow direction and flow accumulation grids
- stream network is defined as cells whose flow accumulation exceeds a **threshold**
- Watershed **outlet** is the cell with highest flow accumulation

Flow Direction Creation



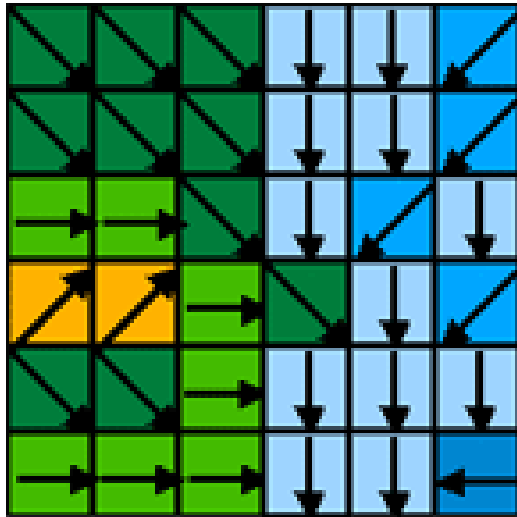
Flow Direction

- If all neighbors are higher than the processing cell, the processing cell is a sink and has an undefined flow direction.
- Cells with undefined flow direction can be flagged. (as sinks using the Sink function in ArcGIS)

Flow Accumulation

- The Flow Accumulation function calculates accumulated flow as the accumulated weight of all cells flowing into each downslope cell in the output raster.
- If no weight raster is provided, a weight of one is applied to each cell, and the value of cells in the output raster will be the number of cells that flow into each cell.
- Cells with a high flow accumulation are areas of concentrated flow and may be used to identify stream channels

Flow Accumulation



0	0	0	0	0	0
0	1	1	2	2	0
0	3	7	5	4	0
0	0	0	20	0	1
0	0	0	1	24	0
0	2	4	7	35	2

32	64	128
16		1
8	4	2

Direction Coding

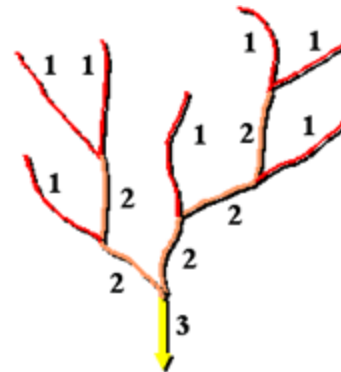
Stream Order Classification

- Stream networks can be delineated from a digital elevation model (DEM) using the output from the Flow Accumulation function. Flow accumulation in its simplest form is the number of upslope cells that flow into each cell.

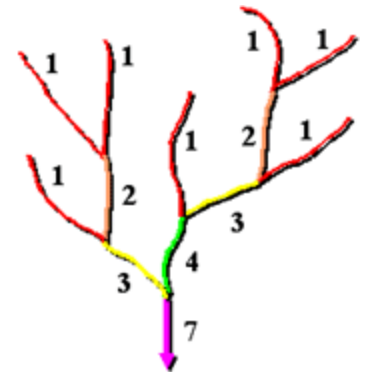
Classification

- Stream ordering is a method of assigning a numeric order to links in a stream network. This order is a method for identifying and classifying types of streams based on their number of tributaries. Some characteristics of streams can be inferred by simply knowing their order.
- Strahler (Horton's)
- Shreve

- Orders



Strahler

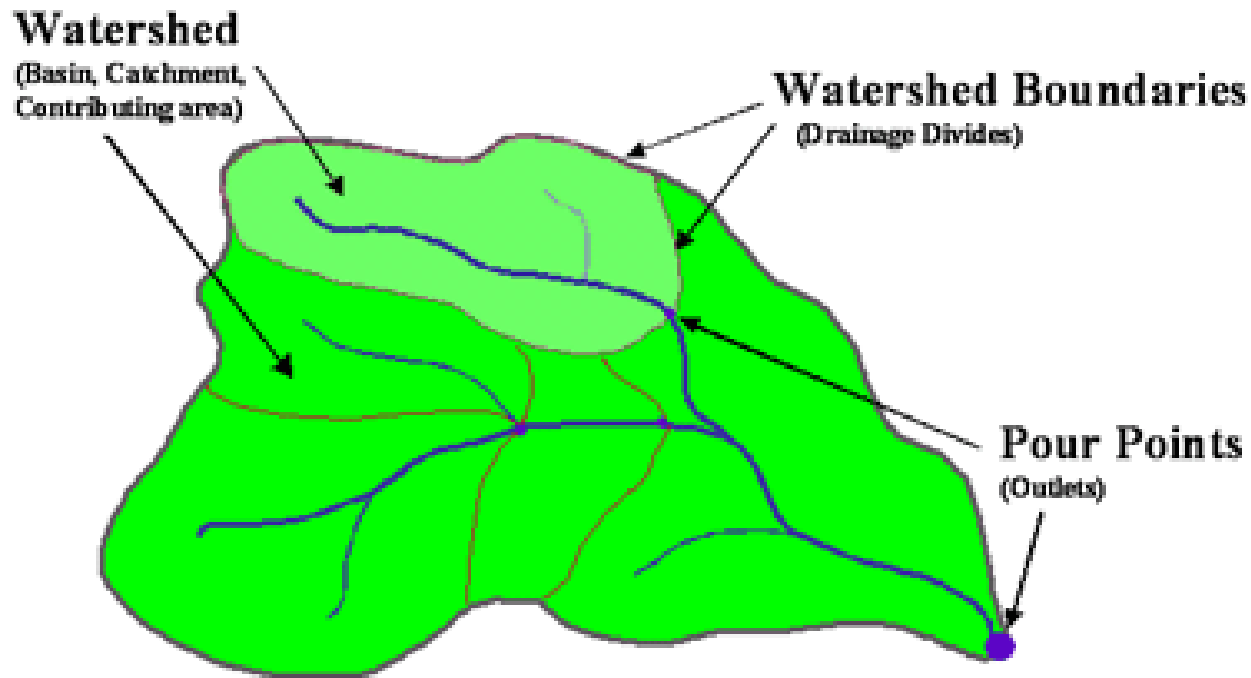


Shreve

DEM applications

- Watershed Delineation
- Parameters for Hydrologic Models
- Volume of Proposed Reservoirs
- Material to be removed for Strip Mining
- Landslide Probability

Characteristics of Watershed



Watershed

- A drainage basin is an area that drains water and other substances to a common outlet. Other common terms for a drainage basin are watershed, basin, catchment, or contributing area.
- This area is normally defined as the total area flowing to a given outlet, or pour point. A pour point is the point at which water flows out of an area.
- This is usually the lowest point along the boundary of the drainage basin. The boundary between two basins is referred to as a drainage divide or watershed boundary.

GIS related Terminology

- The network through which water travels to the outlet can be visualized as a tree, with the base of the tree being the outlet.
- The branches of the tree are stream channels. The intersection of two stream channels is referred to as a node or junction.
- The sections of a stream channel connecting two successive junctions or a junction and the outlet are referred to as stream links.

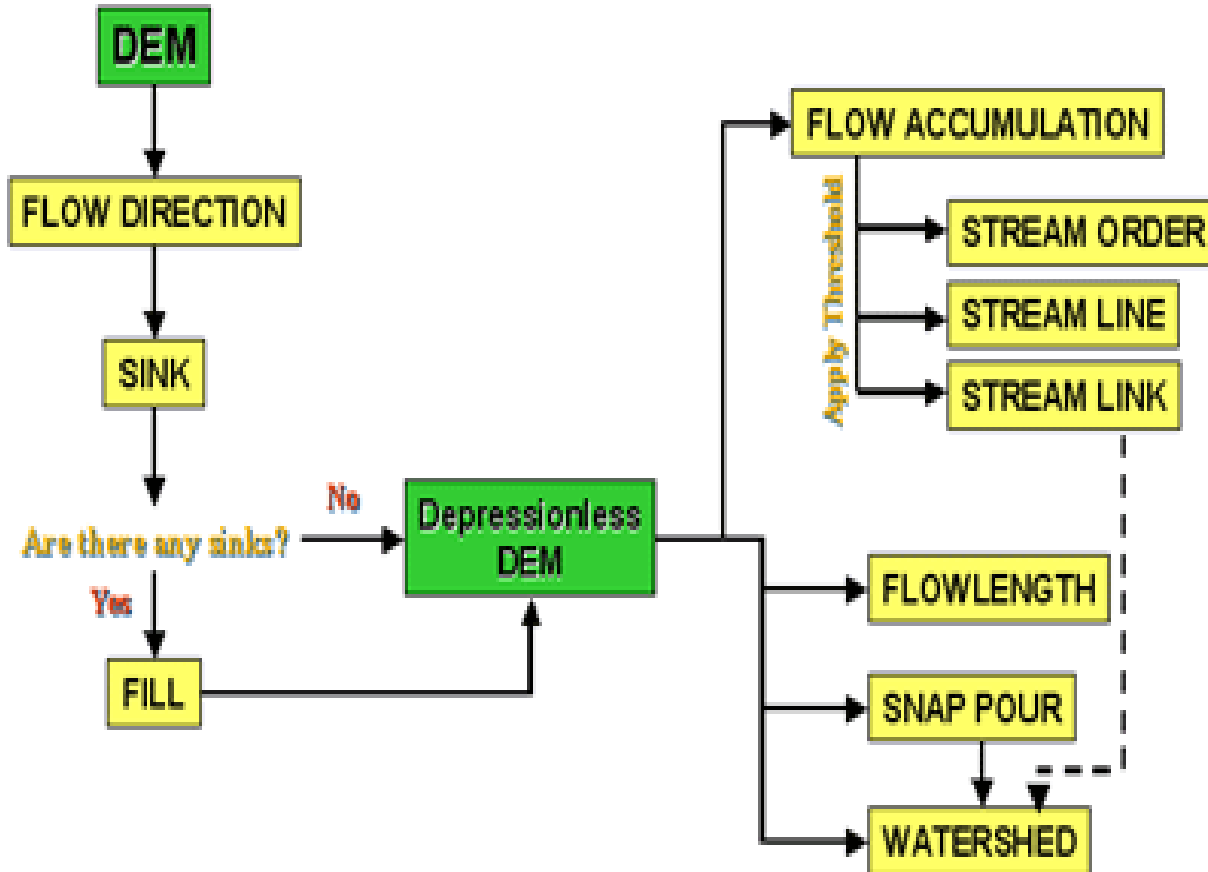
Watershed Delineation

- D-8 Algorithm
- D-8 pour point Algorithm is most commonly used algorithm for delineating stream network and watershed from DEM
- Identify the steepest slope (between each cell and its nearest eight neighboring cells)
- Find the direction of streams
- Calculate stream accumulations

Watershed Delineation



Steps in Watershed Delineation



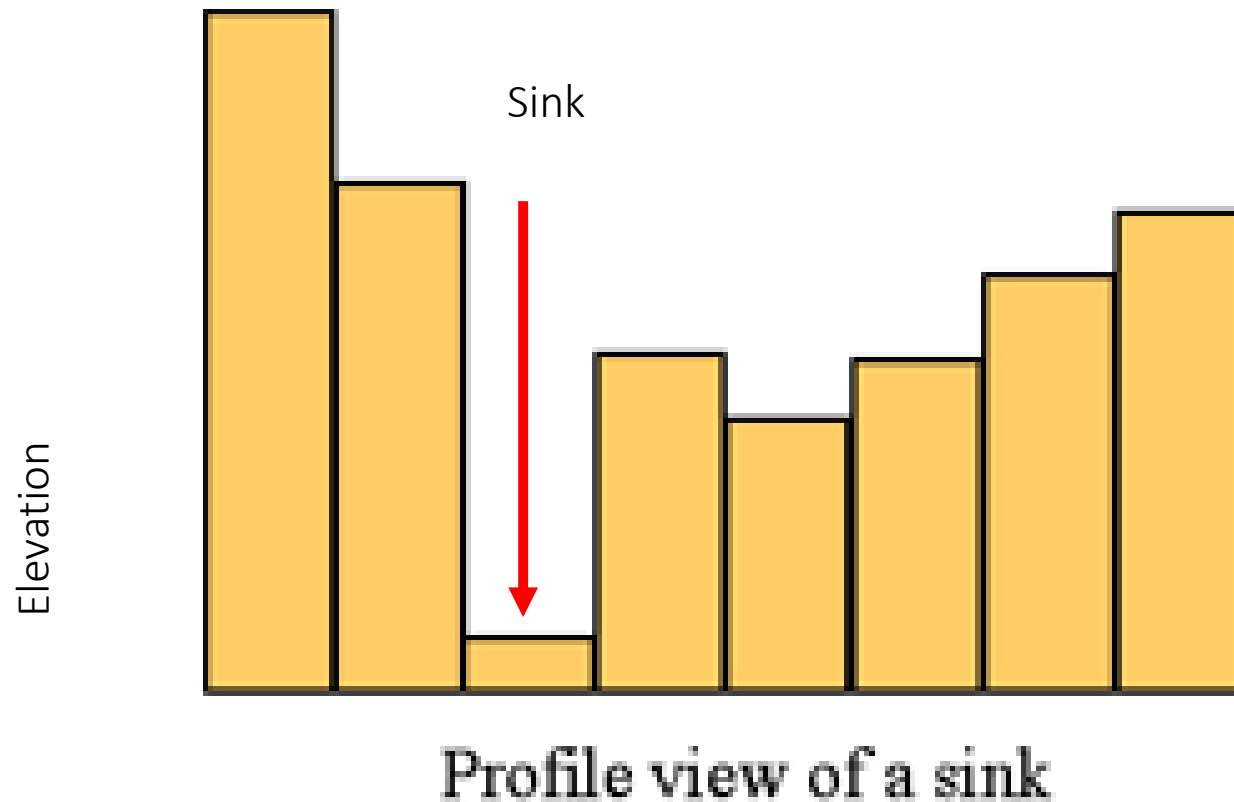
Functions available for Delineation – ArcGIS

Tool	Description	ArcToolbox
Basin	Creates a raster delineating all drainage basins within the Analysis window.	X
Fill	Fills sinks in a surface raster to remove small imperfections in the data.	X
Flow Accumulation	Creates a raster of accumulated flow to each cell by accumulating the weight for all cells that flow into each downslope cell.	X
Flow Direction	Creates a grid of flow direction from each cell to its steepest downslope neighbor.	X
Flow Length	Calculates upstream or downstream distance along a flow path for each cell.	X
Sink	Creates a grid identifying all sinks or areas of internal drainage.	X
Snap Pour Point	Snaps selected pour points to the cell of highest flow accumulation within a specified neighborhood.	X
SnapPour	Snaps selected pour points to the cell of highest flow accumulation within a specified neighborhood.	
Stream Link	Assigns unique values to sections of a raster linear network between intersections.	X
Stream Order	Assigns a numeric order to segments of a grid representing branches of a linear network.	X
Stream To Feature	Converts a raster representing a raster linear network to a feature class.	X
StreamShape	Converts a grid representing a raster linear network to a shapefile.	
Watershed	Determines the contributing area above a set of cells in a grid.	X

Depression Less DEM

- A digital elevation model (DEM) free of sinks
 - A depressionless DEM—is the desired input to the flow direction process.
 - The presence of sinks may result in an erroneous flow–direction raster.
 - In some cases, there may be legitimate sinks in the region (e.g. quarries, natural depressions, etc)
 - It is important to understand the morphology of the area well enough to know what features may truly be sinks on the surface of the earth and which are merely errors in the data.

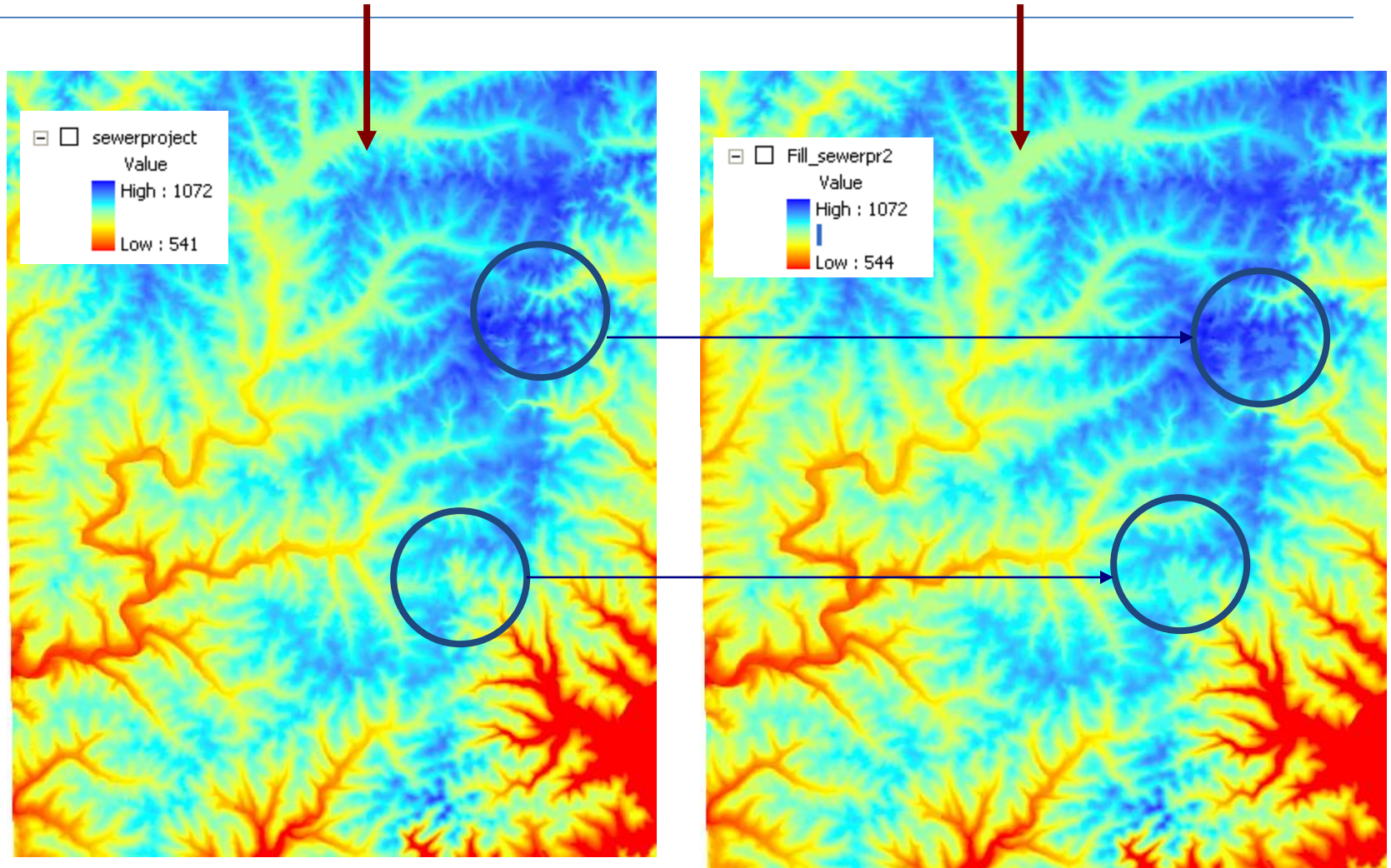
View of Sink



Filling Sinks

- Identifying sinks Sinks can be located using the **Sink function**.
- This function requires a direction raster that is created by the Flow Direction function.
- The result is a raster that identifies any existing sinks in the data.
- Depending on the results, you can fill the sinks, or you can use the output to help determine the fill limit.
- Sinks can be filled using the Fill function.

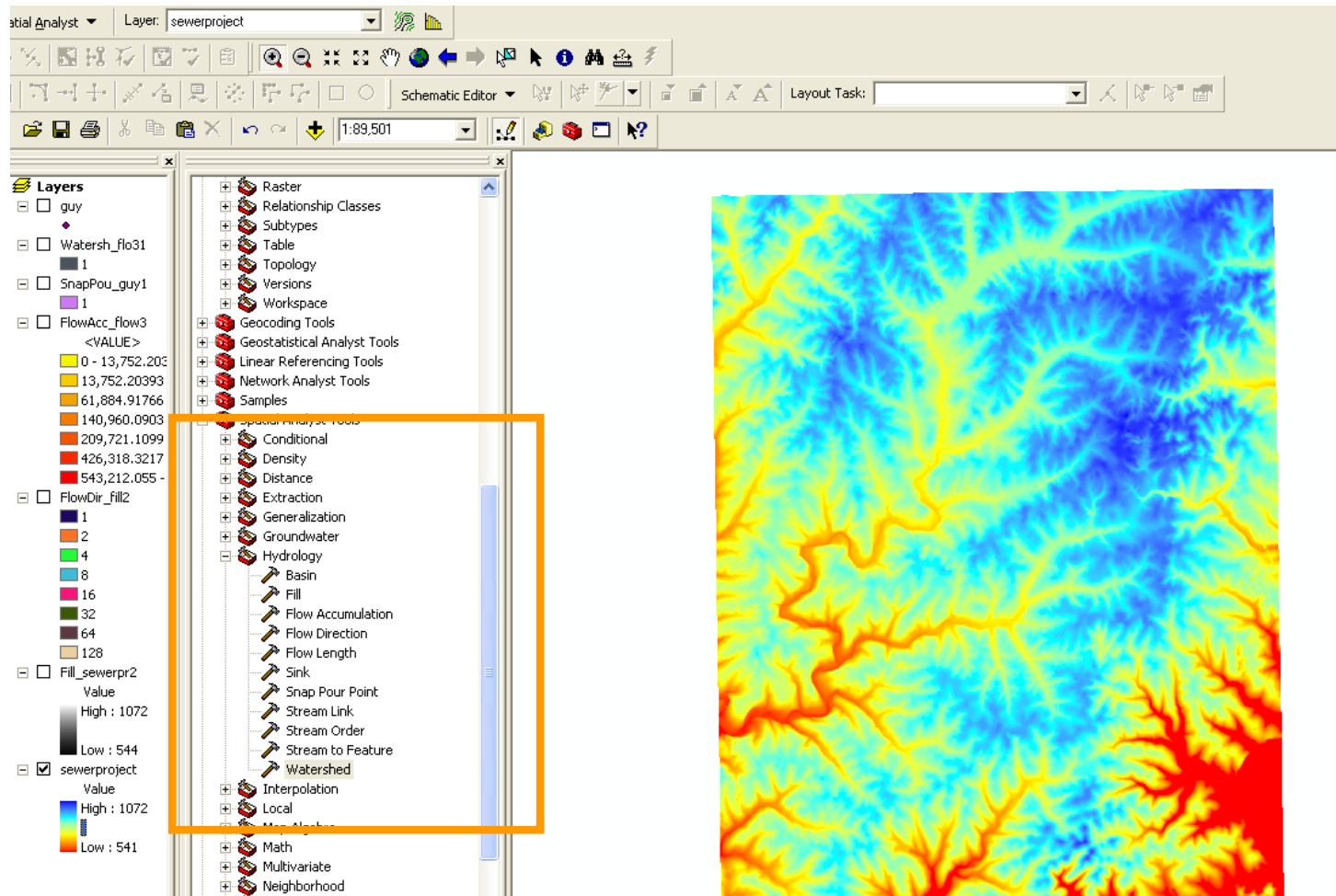
Before and 'After Fill'



Steps and Creation of delineated Watershed

- Fill (fill the pits)
- Flow Direction Calculations
- Flow Accumulation
- Create a shape file with Pour Points
- Snap the pour Points
- Delineate Watershed
- Change (Watershed – from Raster form) to feature form (vector form)

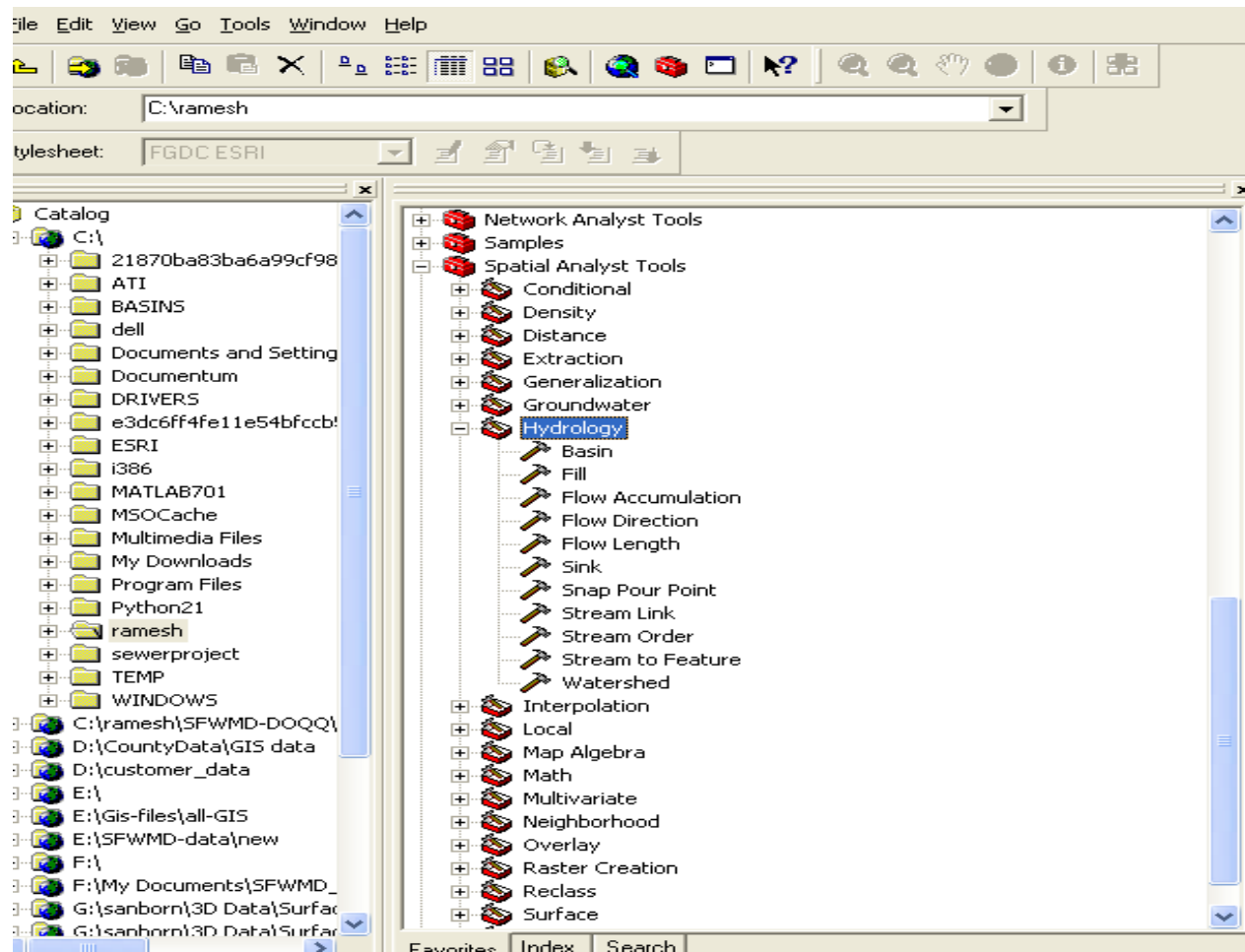
DEM – Hydrology Tools



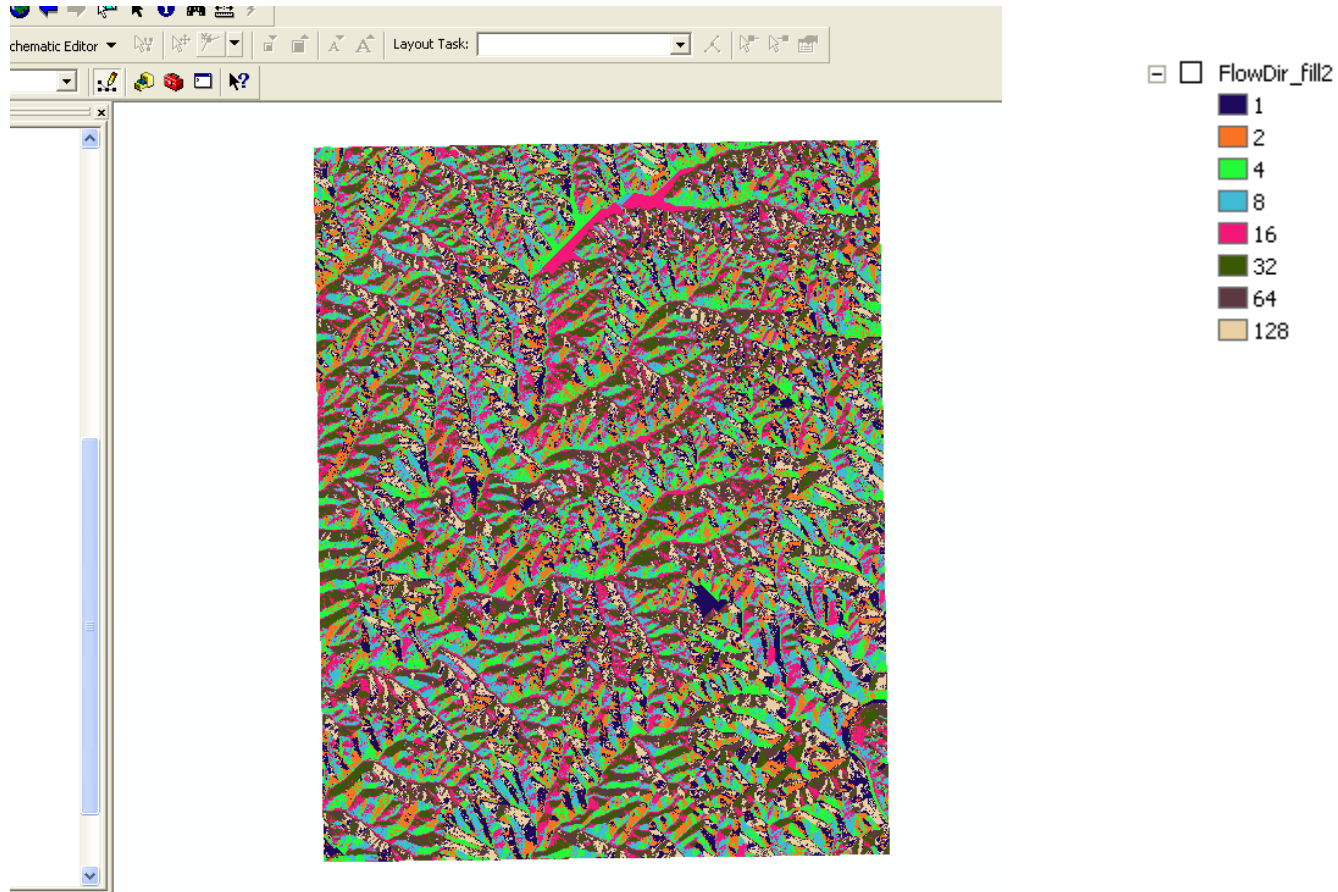
Digital Watershed Delineation

- Load ArcMap
- Define properties of Layers
- Load DEM
- Start ArcTool Box
- Select Spatial Analyst
- Select “Hydrology” Sub Menu under Spatial Analyst.

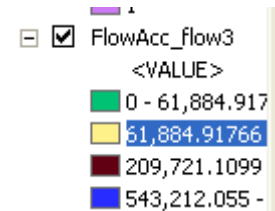
Spatial Analyst



Flow Direction

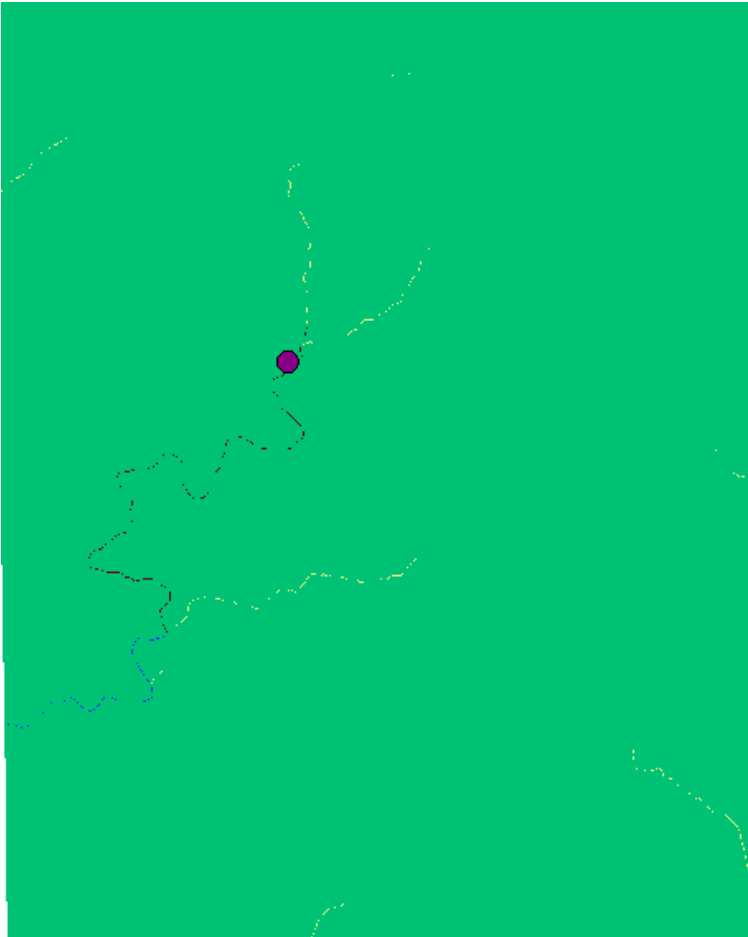


Flow Accumulation



Darker Areas
Represent higher
Flow accumulations

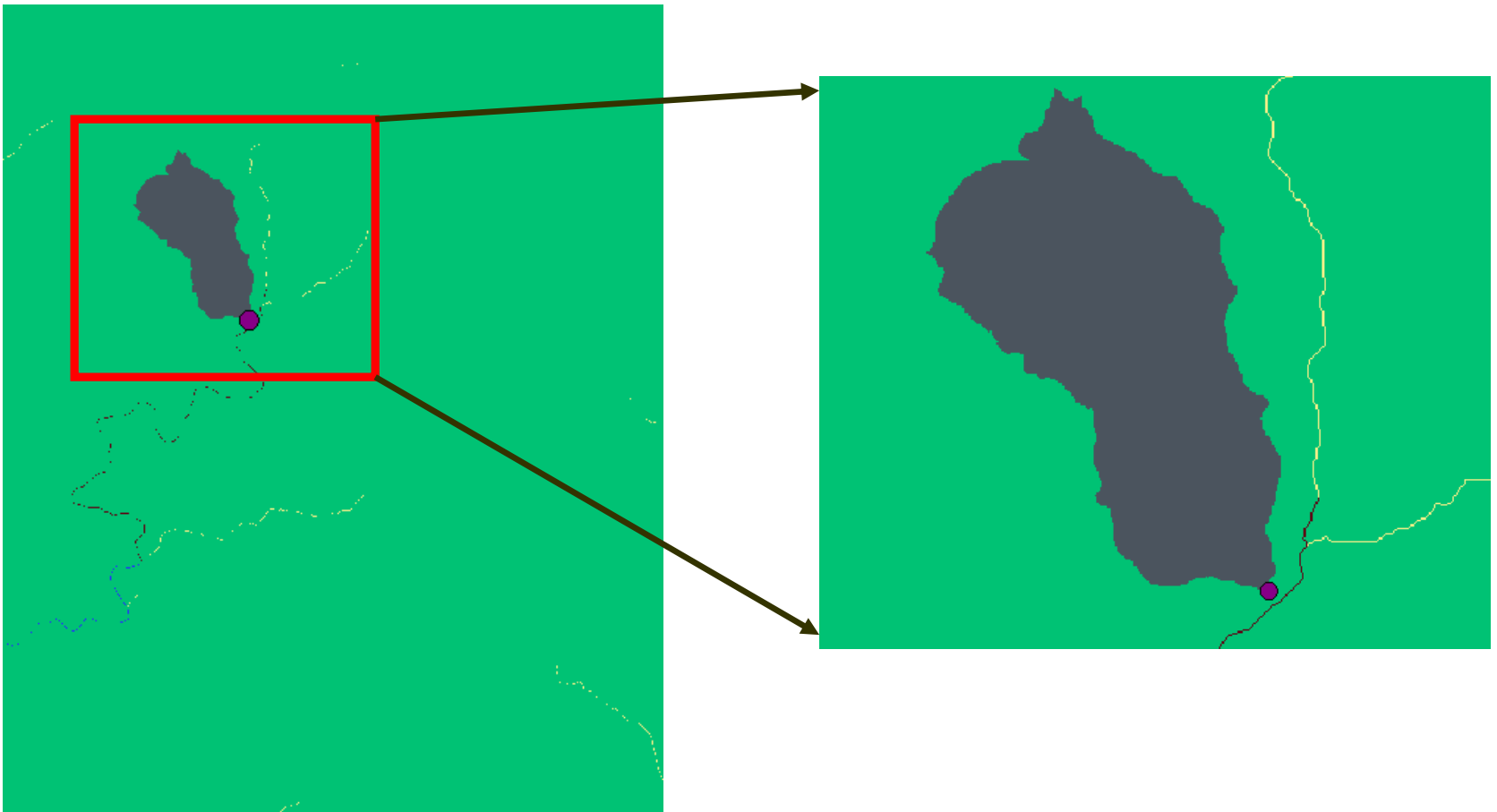
Create Pour point (outlet)



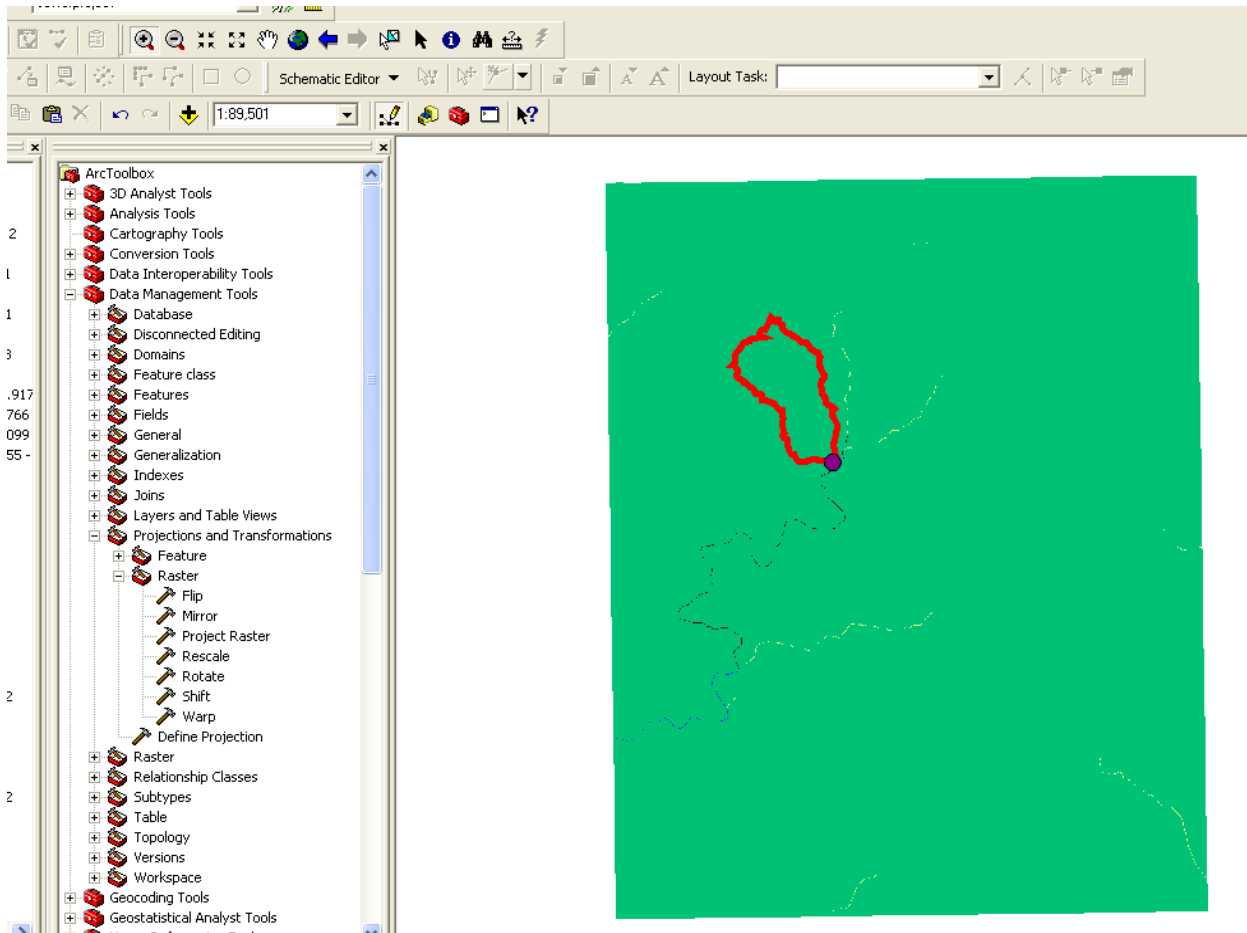
Pour Point (outlet)

To create a pour point use
“ArcCatalog” to create a
shapefile (point)

Delineated Watershed



Vector form of Watershed



Start Spatial
Analyst

Use Convert
option to
create –vector
form of
watershed.

Use Raster to
feature

Fill the Pits

- DEM creation results in **artificial pits** in the landscape
- A pit is a set of one or more cells which has **no downstream cells** around it
- Unless these pits are filled they become **sinks** and isolate portions of the watershed
- **Pit filling** is the first step in correcting the problems with DEM

FILL Approach

- Assumes that all sinks are caused by underestimated elevation values
- Also Sinks can also be caused by overestimated values.

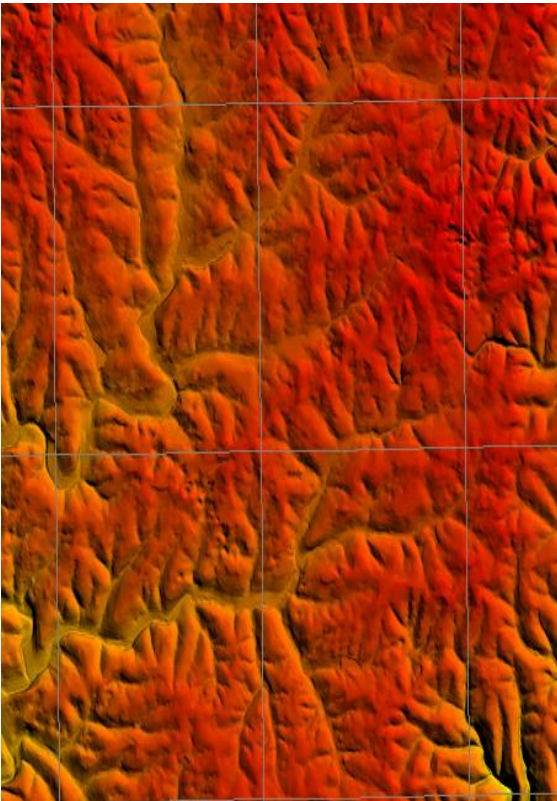
Stream Burning



- DEM-based stream delineation may not be accurate in flat areas or if the DEM resolution failed to capture the important topographic information
- Problem can be solved by a process referred to as “stream burning”
- Using known stream locations, streams are burned into the DEM layer.

Stream Burning

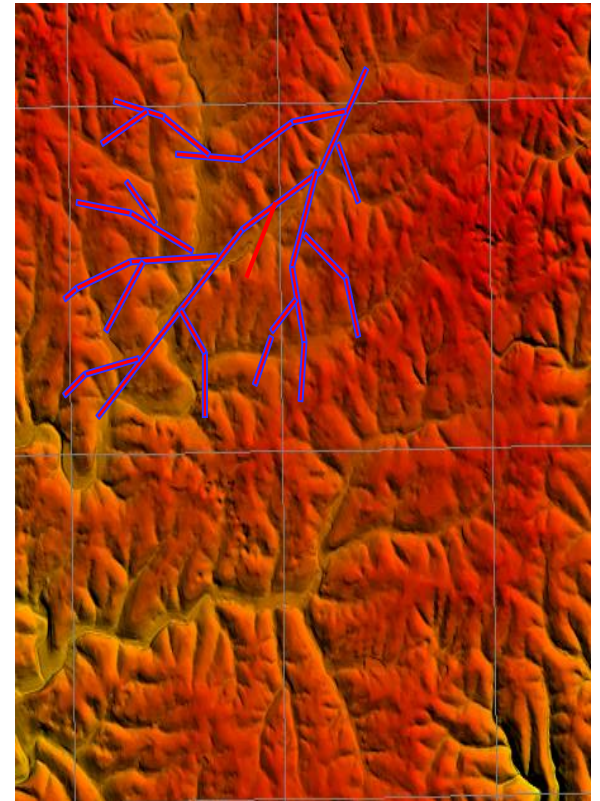
DEM



Stream Layer



Burned DEM



Stream Burning

- This Process modifies the DEM so that the flow of water is forced into the know stream locations.
- The cell elevations where streams are located are lowered (artificially) to facilitate this process or surrounding cells are elevated.
- The phrase “ Burning in” indicates that the streams have been forced or burned into the DEM.
- This methodology should be used with caution.

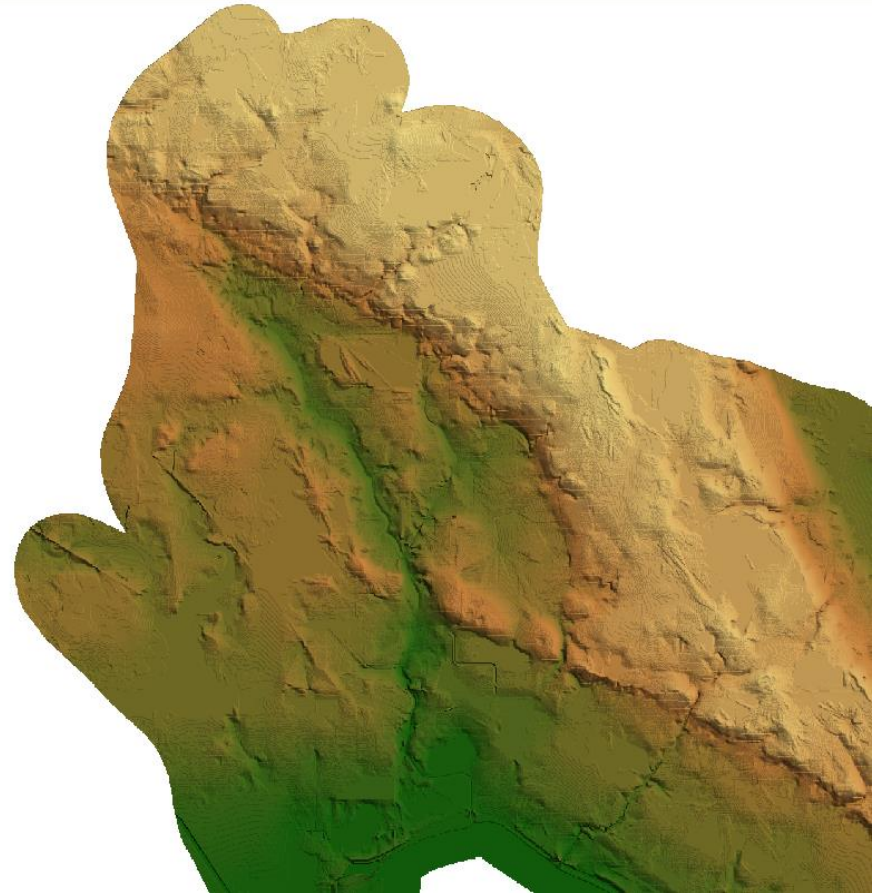
DEM resolution

- The resolution of DEM will affect the average slope of the watershed, drainage length and direction and streams.
- The main objective of this case study is to assess the effect of DEM resolution on hydrological and water quality modeling.

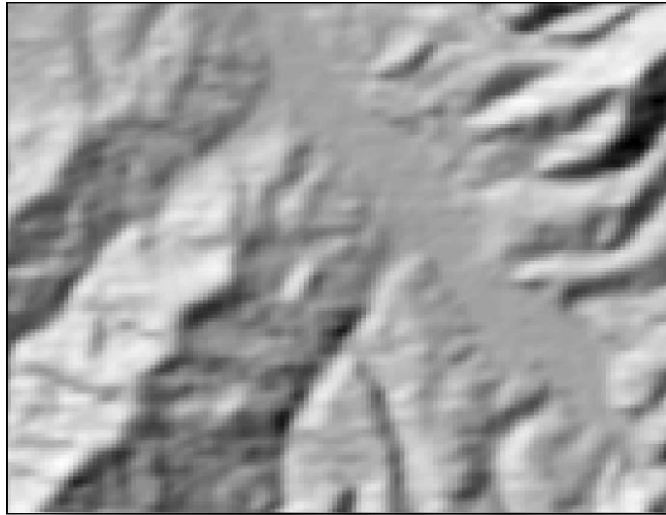
100 m X 100 m

DEMs

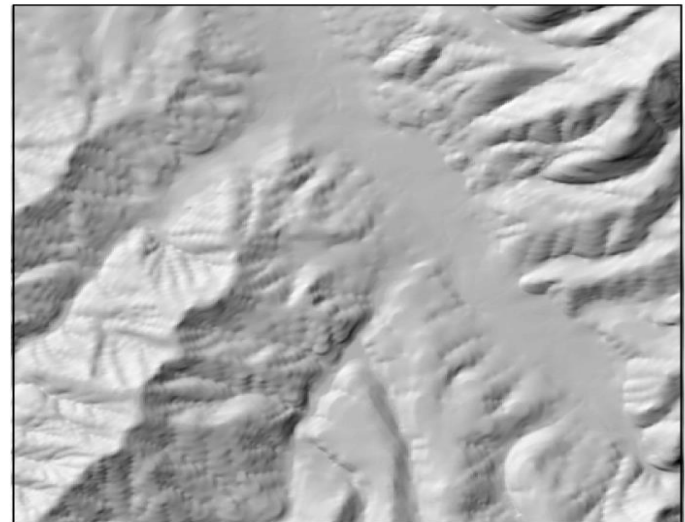
30 m X 30 m



DEMs at Different Resolutions



(a)



(b)

DEMs at a 30-meter resolution (a) and a 10-meter resolution (b).

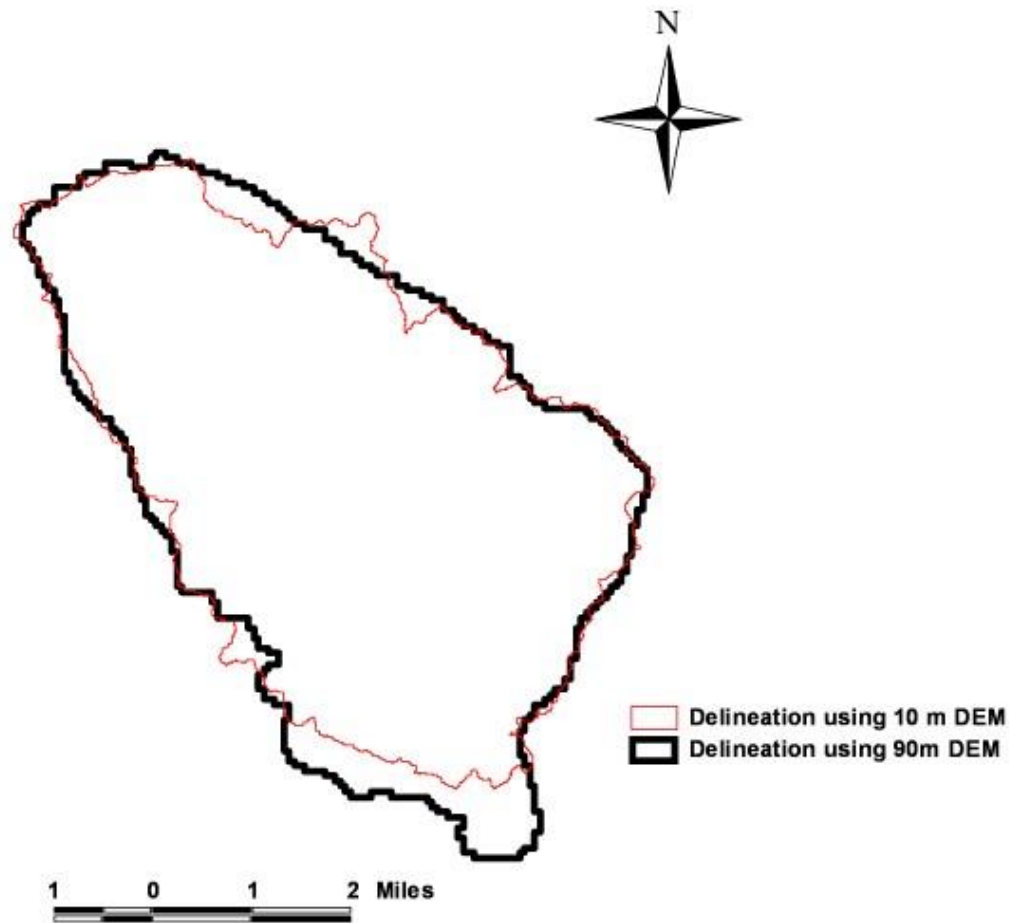
Effect of DEM resolution

Watershed Characteristics	DEM resolution	
	10 m x 10 m	90 m x 90 m
Area (acres)	14005	14992
Slope (ft/ft)	0.0115	0.0118
Minimum Elevation (ft)	256	272
Maximum Elevation (ft)	320	328

Effect of DEM – Land USe

Major Land Use Type	DEM Resolution	
	10 m x 10 m	90 m x 90 m
	Area (acres)	
Urban or Built-up land	3254	3254
Agricultural Land	10519	11366
Forest Land	72	72
Barren Land	162	160
Total	14007	14852

Watershed Areas using two different DEMs



DEM resolution

- The resolution of digital elevation model (DEM) used for watershed delineation and data extracted from the delineated watershed relevant to hydrography plays an important role in the calibration of hydrological models.
- Furthermore, the water quality modeling of streams depends on the hydrologic features of the watershed that are obtained using DEMs. DEM resolution can affect hydrologic modeling parameters, water quality simulation results.
- Resolution will influence water quality management strategies (e.g. total maximum daily load development) which quantify pollutant load reductions linked to non-point pollution sources within a watershed.
- The watershed area considered based on a specific DEM resolution will alter the land use composition and thereby ultimately affecting the non-point source loadings in the watersheds.

Hydrological Modeling Results (example)

Errors (Simulated-Observed)	DEM resolution		
	10 m x 10 m	90 m x 90 m	90m x 90 m*
Error in total volume	-1.00	8.65	4.68
Error in 50% lowest flows	-10.36	-6.72	0.91
Error in 10% highest flows	-6.75	7.60	-6.38
Seasonal volume error – Summer	19.96	33.27	29.08
Seasonal volume error – Fall	17.39	40.60	23.58
Seasonal volume error – Winter	-13.38	-3.42	-9.06
Seasonal volume error – Spring	1.36	2.87	7.55
Error in storm volumes	25.48	50.86	18.90
Error in summer storm volumes	47.97	69.27	53.23

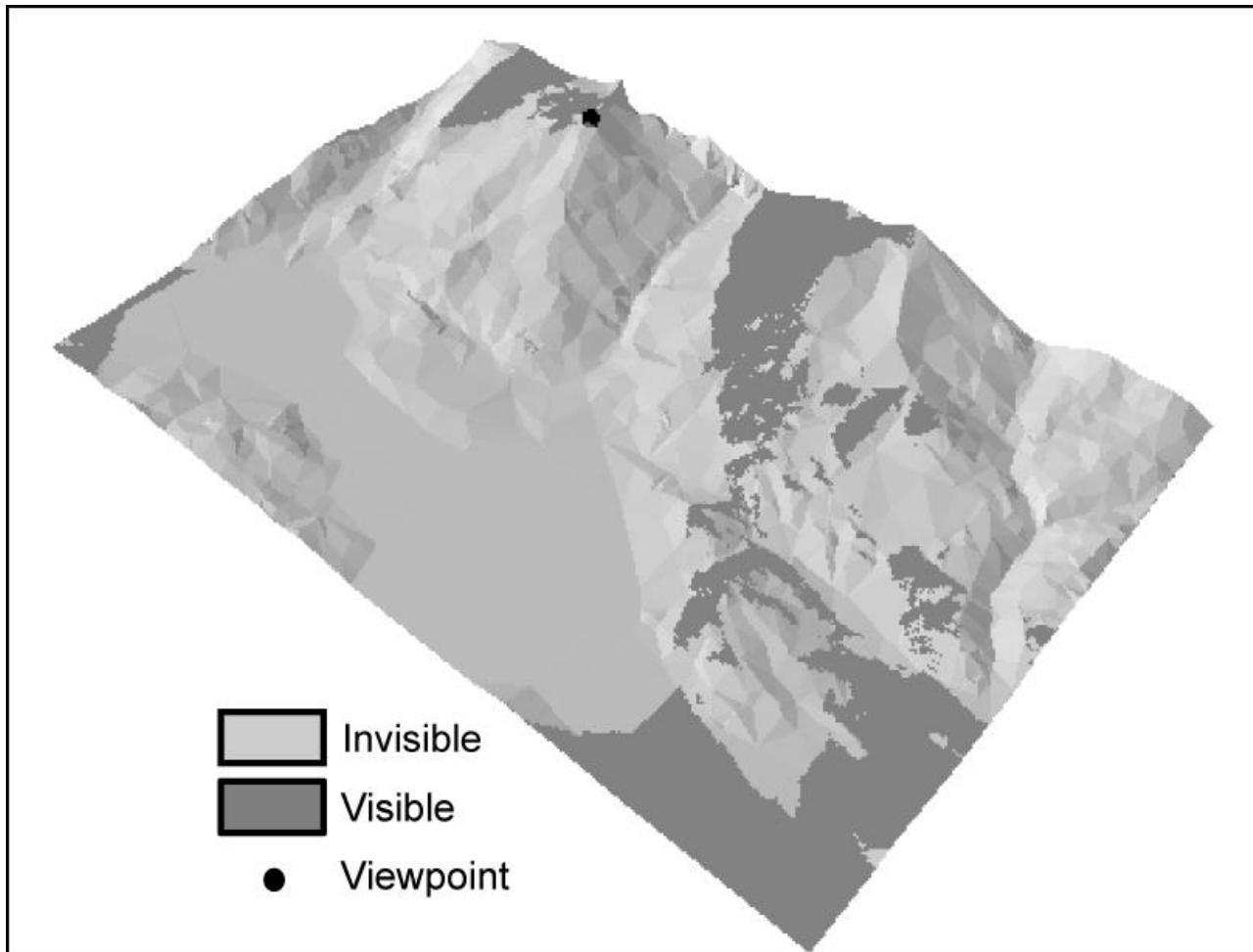
*results based on calibration parameters of 10 m x 10 m DEM

Viewsheds

Viewshed Analysis

A viewshed refers to the portion of the land surface that is visible from one or more viewpoints. The process for deriving viewsheds is called viewshed or visibility analysis.

Viewshed Example



Generation of Datasets

- Generation of gridded datasets for hydrological modeling
- Distributed hydrological modeling studies require data spatially varying processes/parameters
 - Land cover
 - Precipitation
 - Temperature
 - Evapotranspiration
 - Soils
 - others

Data

- Gridded data at a fixed spatial resolution (tessellation) can be used for modeling purposes.
- However, in many instances these data is not readily available.
- If only point data is available, spatial interpolation is used to develop a surface and then the surface is used to develop gridded dataset.

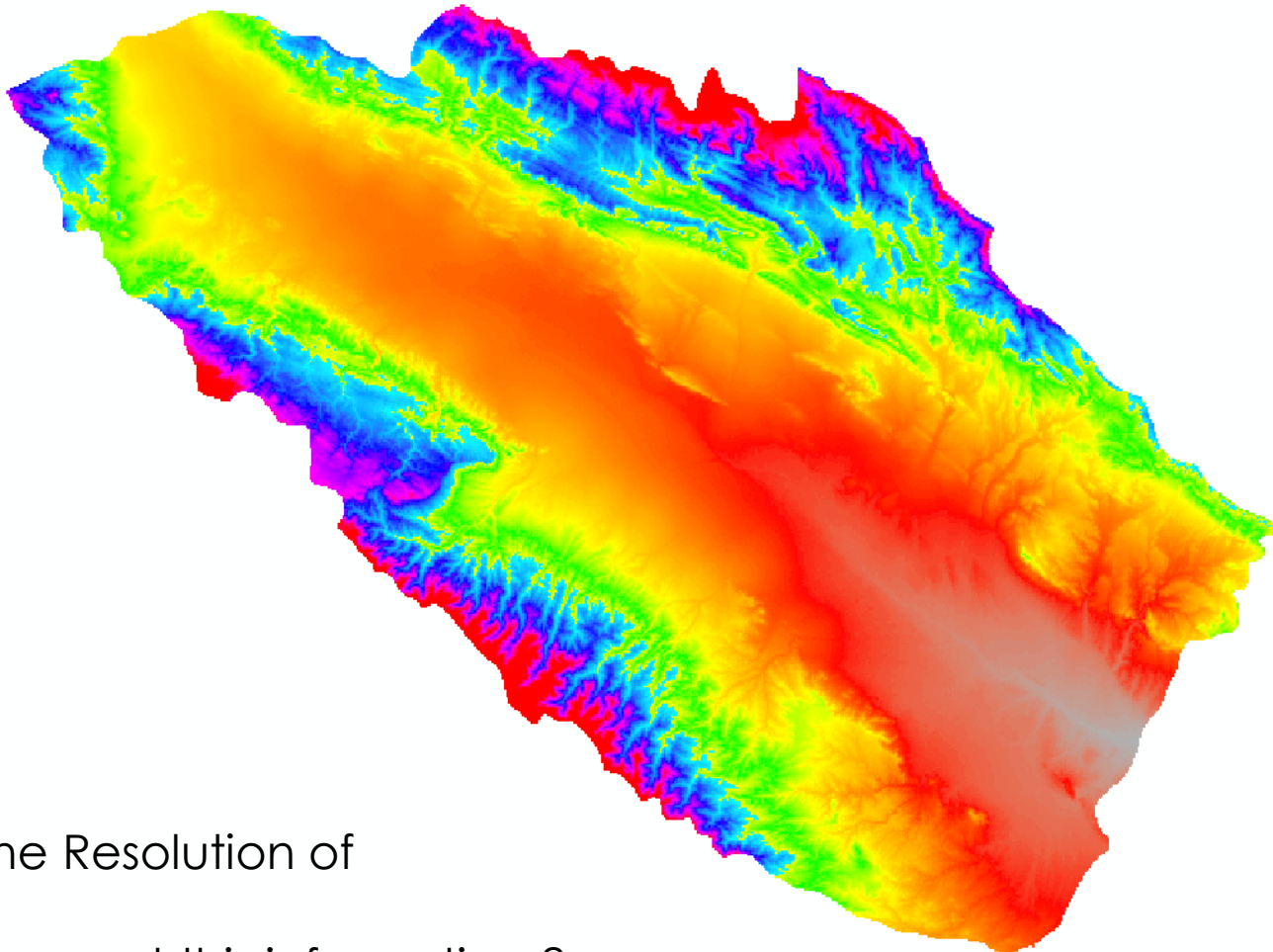
Point Data to Gridded Data

- Obtain point observation data
- Obtain spatial grid (of a specific resolution). The obtained grid can be a Digital Elevation Model (DEM) or any grid with no attributes attached to it.
- Use any spatial analysis software to interpolate spatial data to develop a surface.
 - Spatial interpolation requires selection of a specific interpolation technique that is appropriate for process under consideration.
 - Spatial interpolation method needs to be evaluated by selection of calibration and validation points.
 - K fold validation is also recommended to obtain confidence in the spatially interpolated estimates
 - Multiple performance measures can be evaluated before selecting a method.

Steps

- Available data
 - DEM data from a region in Iran.
 - DEM resolution (250 x 250 meters)

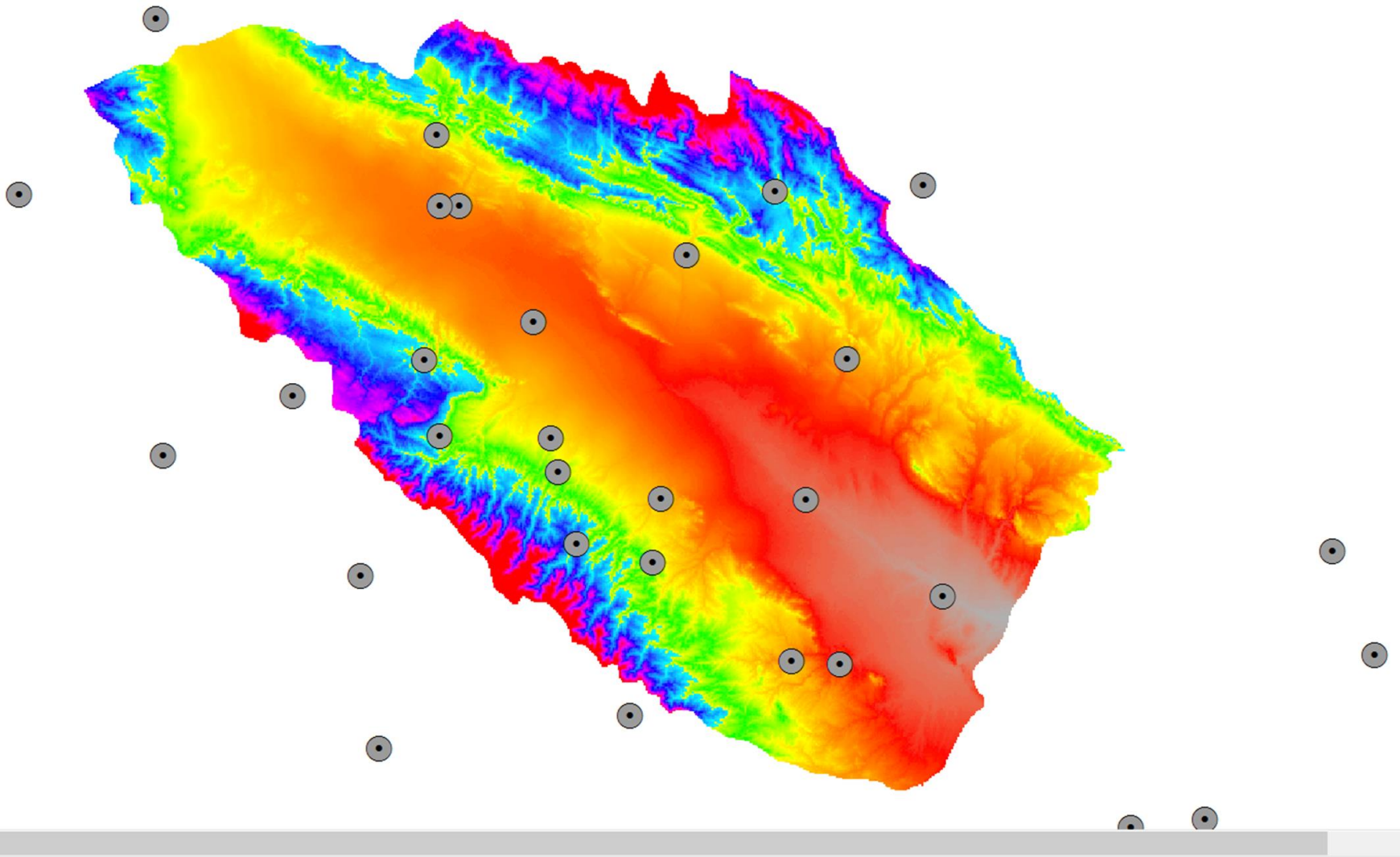
Digital Elevation Model (DEM)



What is the Resolution of
DEM ?

How do you get this information ?

Area with rain gage stations



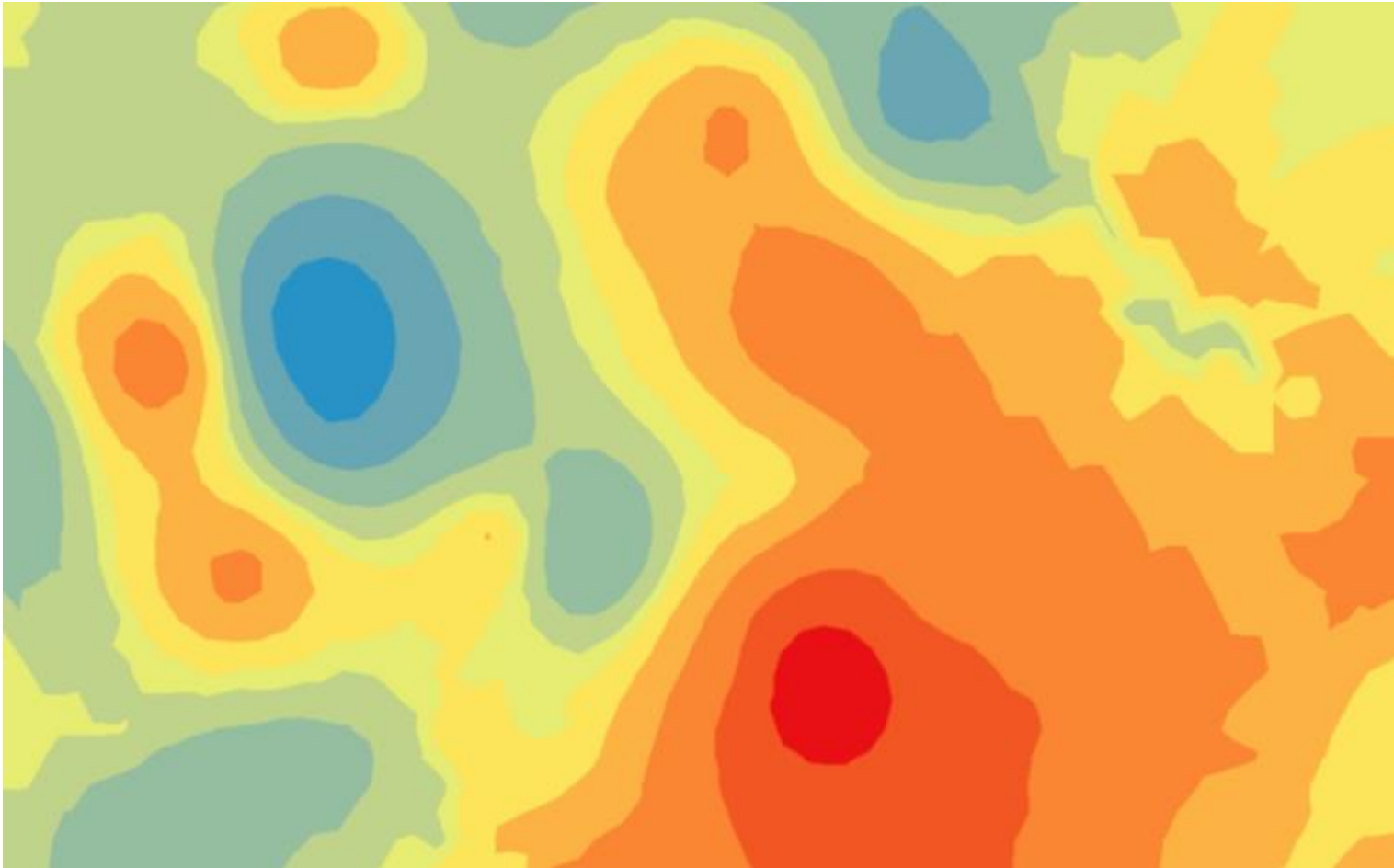
Steps

- Interpolate using any spatial interpolation technique.
- Convert the interpolated surface to a raster.
- When converting to raster make sure the environmental settings (export settings) are correctly indicated.
 - Define the output coordinate system (same as the base layer [i.e. DEM layer in this case]).
 - Define the processing extent. In this case you want the extent to be the same as that of base year (i.e. DEM layer).

Steps

- In Raster Analysis processing, select the mask option
 - Select the base layer (i.e. DEM).
- The mask option helps in confining the interpolated surface to the base layer extent.
- The operation when completed will result in a dataset that can be exported to ascii or a text file for further processing of data by a hydrological simulation model.

Spatial Interpolation



Raster Conversion

GA Layer To Grid

Input geostatistical layer
Kriging

Output surface raster
C:\Users\hchen\Documents\ArcGIS\Default.gdb\GALay16

Output cell size (optional)
1028.72

Number of points in the cell (horizontal) (optional)
1

Number of points in the cell (vertical) (optional)
1

GA Layer To Grid

Exports a Geostatistical layer to a raster.

OK Cancel Environments... << Hide Help Tool Help



Environment Settings



- ✧ **Workspace**
- ✧ **Output Coordinates**
- ✧ **Processing Extent**
- ✧ **XY Resolution and Tolerance**
- ✧ **M Values**
- ✧ **Z Values**
- ✧ **Geodatabase**
- ✧ **Geodatabase Advanced**
- ✧ **Fields**
- ✧ **Random Numbers**
- ✧ **Cartography**
- ✧ **Coverage**
- ✧ **Raster Analysis**
- ✧ **Raster Storage**
- ✧ **Geostatistical Analysis**
- ✧ **Parallel Processing**
- ✧ **Terrain Dataset**

Environment Settings

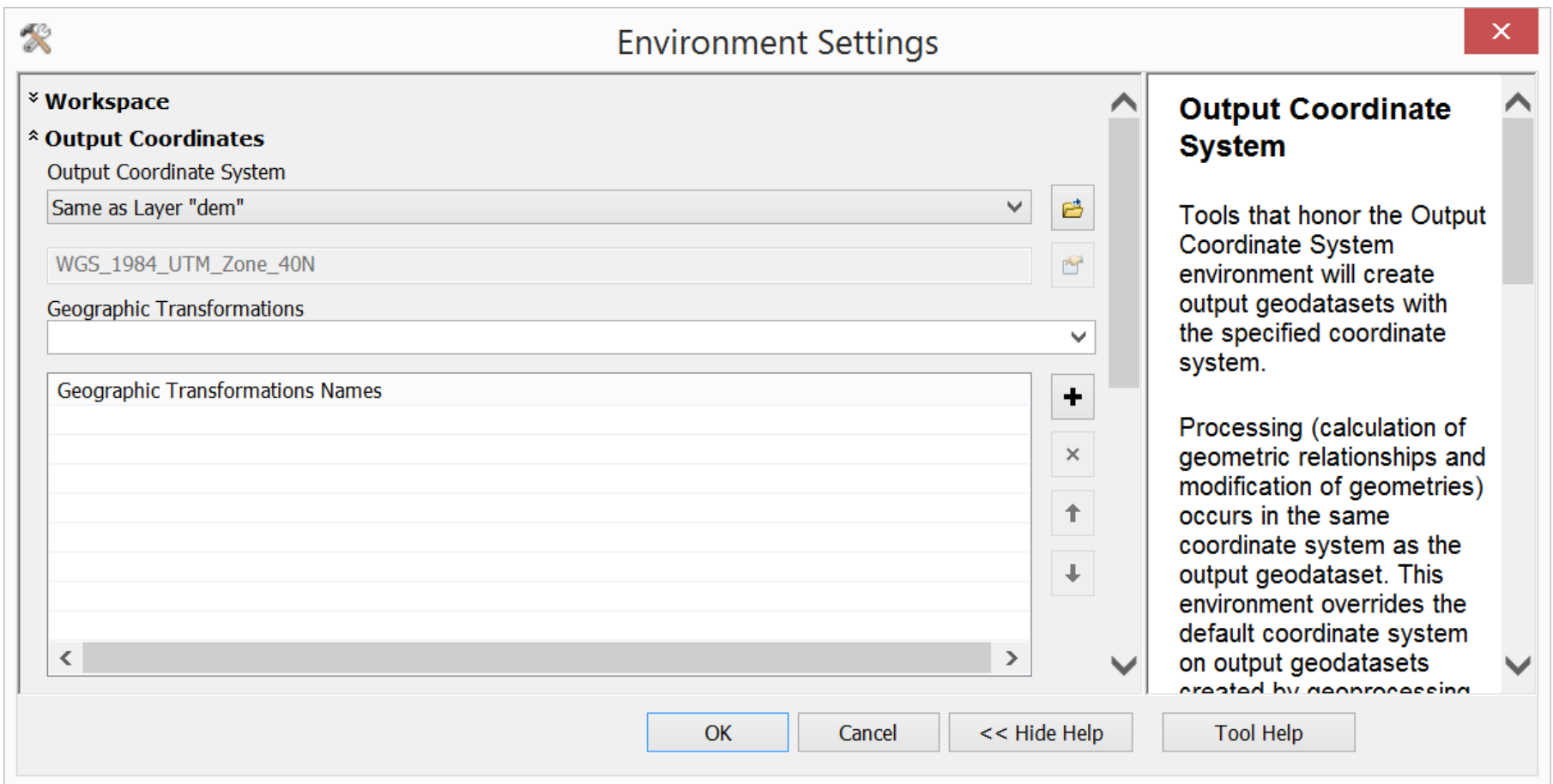
Environment settings specified in this dialog box are values that will be applied to appropriate results from running tools. They can be set hierarchically, meaning that they can be set for the application you are working in, so they apply to all tools; for a model, so they apply to all processes within the model; or for a particular process within a model. Environments set for a process within a model will override all other

OK

Cancel

<< Hide Help

Tool Help



The dialog box is titled "Environment Settings" and has a close button (X) in the top right corner. It is divided into two main sections: "Workspace" and "Output Coordinate System".

Workspace Section:

- Output Coordinates:** A dropdown menu is set to "Same as Layer 'dem'". Below it is a text field containing "WGS_1984_UTM_Zone_40N".
- Geographic Transformations:** A dropdown menu is currently empty. Below it is a list box titled "Geographic Transformations Names" which is empty. To the right of the list box are buttons for adding (+), removing (x), and moving up/down (arrows).

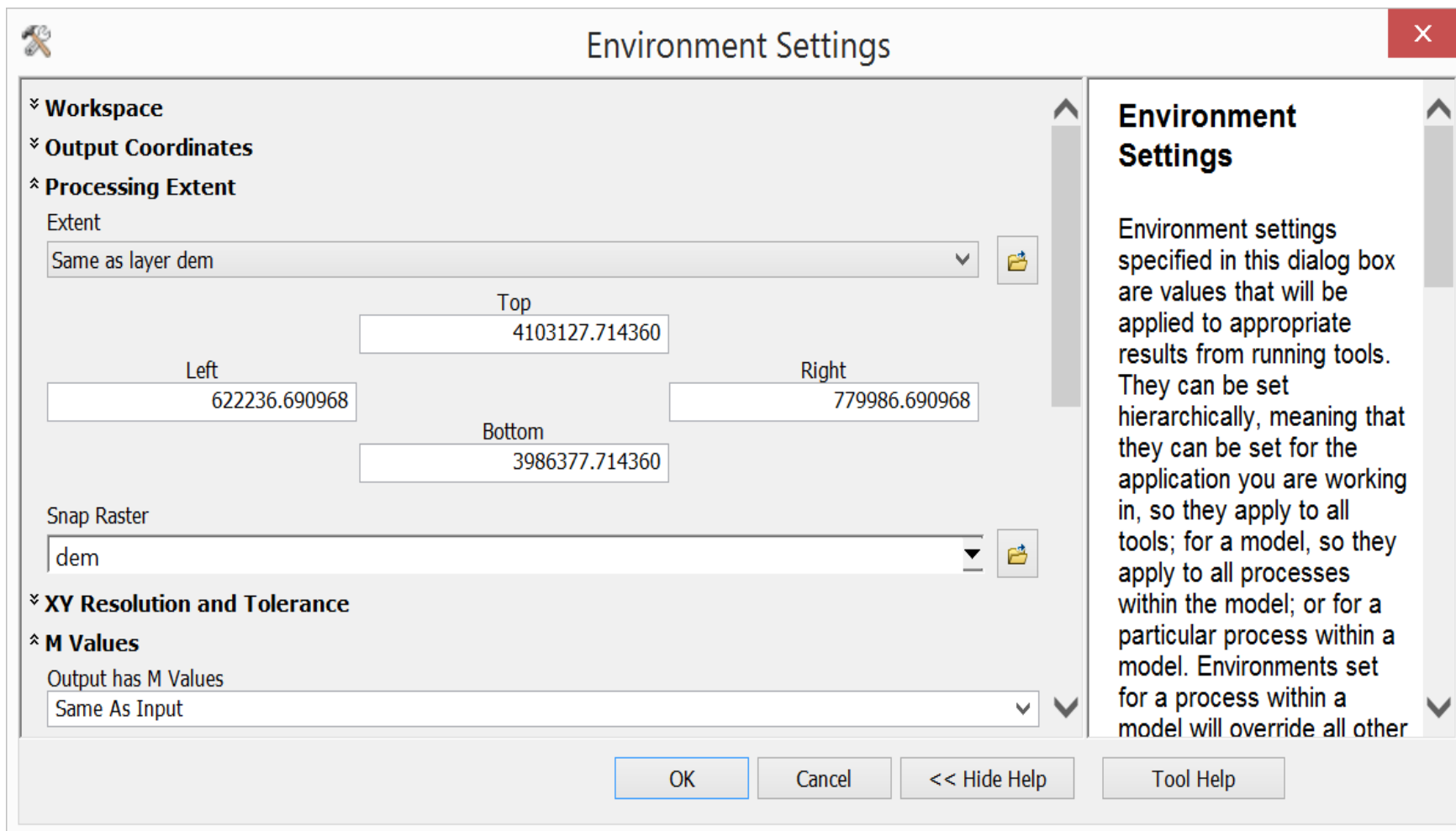
Output Coordinate System Section:

Output Coordinate System

Tools that honor the Output Coordinate System environment will create output geodatasets with the specified coordinate system.

Processing (calculation of geometric relationships and modification of geometries) occurs in the same coordinate system as the output geodataset. This environment overrides the default coordinate system on output geodatasets created by geoprocessing.

Buttons: At the bottom are four buttons: "OK", "Cancel", "<< Hide Help", and "Tool Help".



The image shows a software dialog box titled "Environment Settings". It has a close button (X) in the top right corner. The dialog is divided into two main sections. The left section contains settings for "Workspace", "Output Coordinates", "Processing Extent", "Snap Raster", "XY Resolution and Tolerance", and "M Values". The right section, titled "Environment Settings", contains a descriptive text about the settings. At the bottom, there are four buttons: "OK", "Cancel", "<< Hide Help", and "Tool Help".

Environment Settings

Workspace

Output Coordinates

Processing Extent

Extent

Same as layer dem

Top

4103127.714360

Left

622236.690968

Right

779986.690968

Bottom

3986377.714360

Snap Raster

dem

XY Resolution and Tolerance

M Values

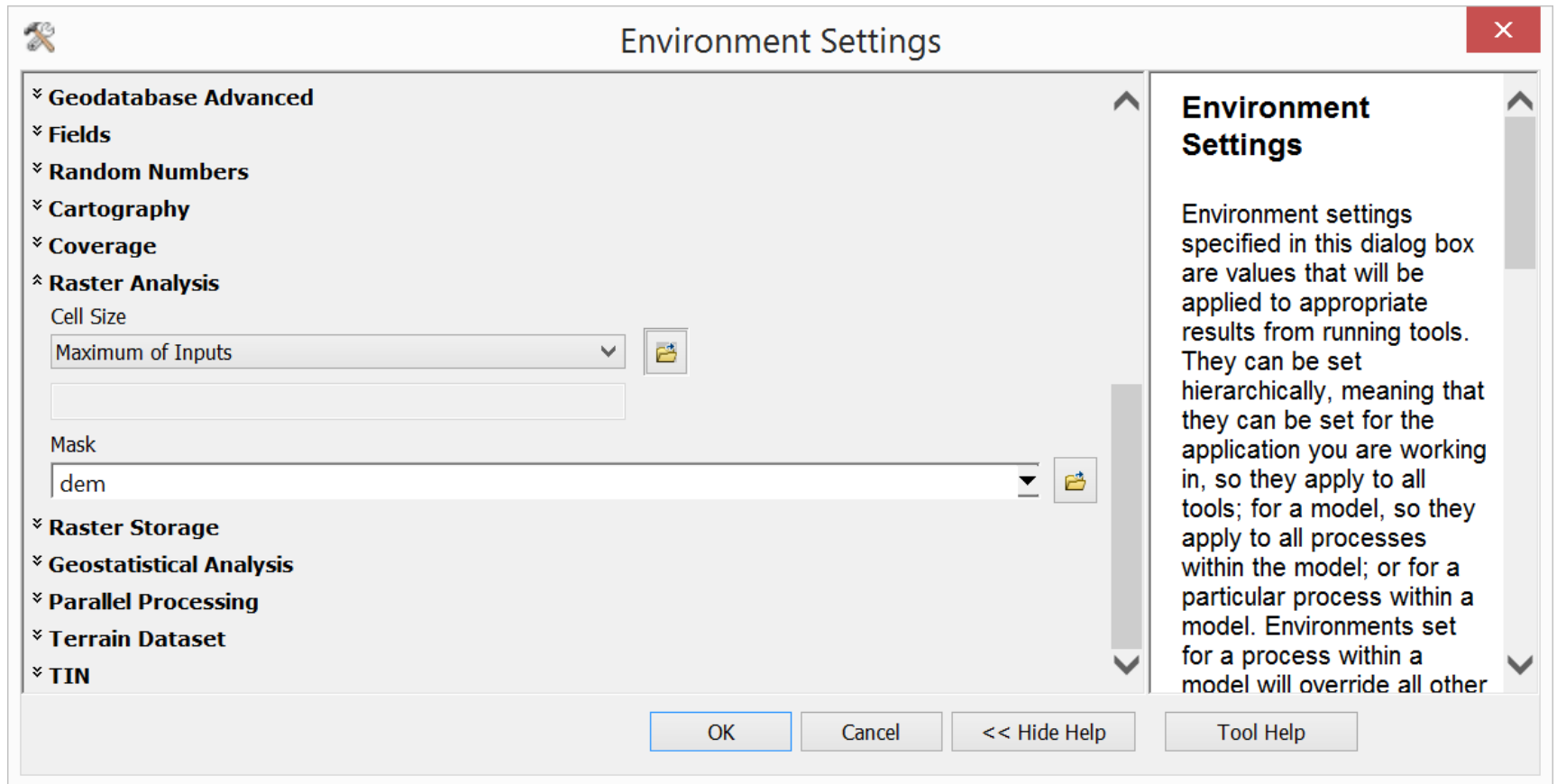
Output has M Values

Same As Input

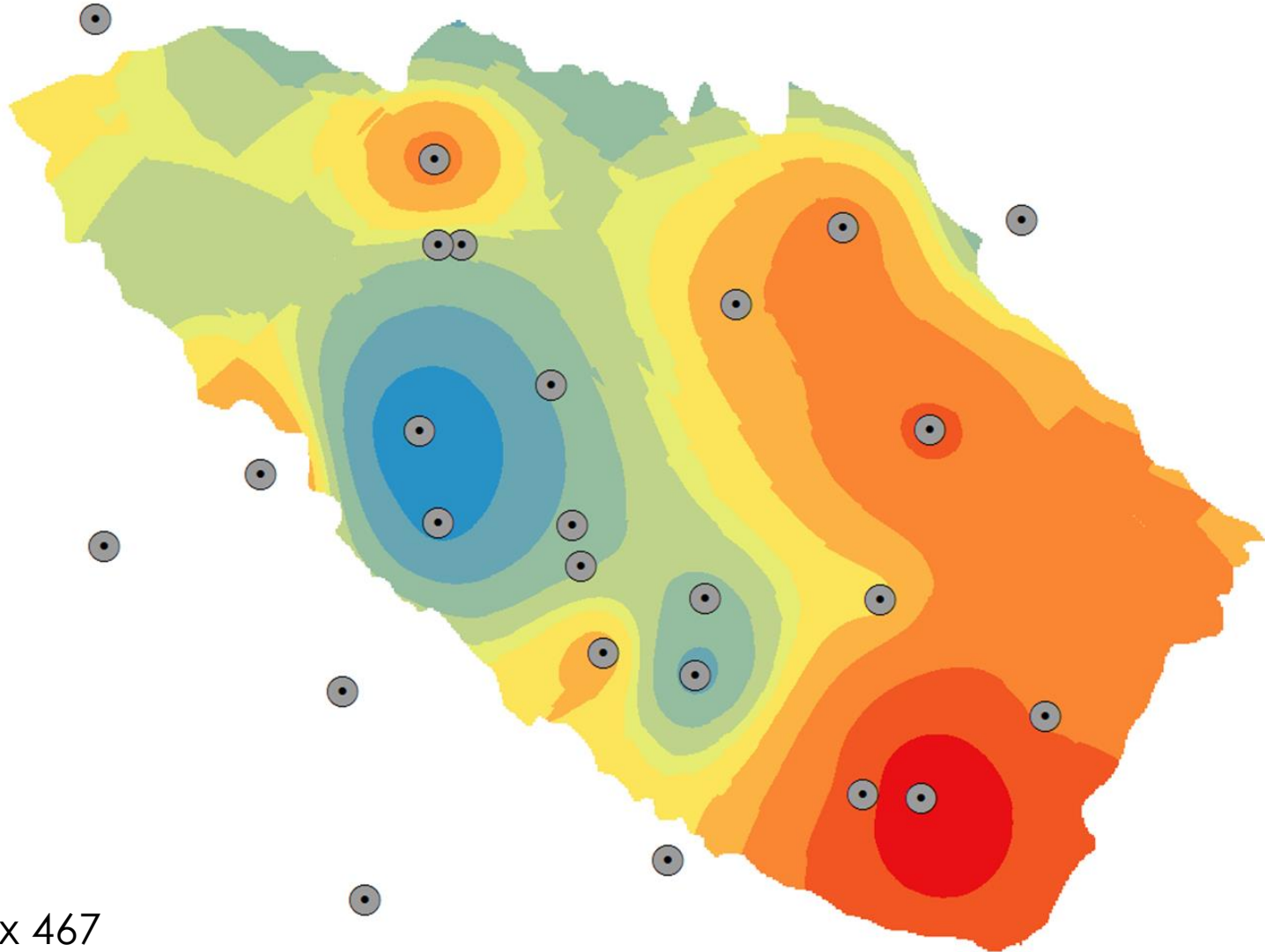
Environment Settings

Environment settings specified in this dialog box are values that will be applied to appropriate results from running tools. They can be set hierarchically, meaning that they can be set for the application you are working in, so they apply to all tools; for a model, so they apply to all processes within the model; or for a particular process within a model. Environments set for a process within a model will override all other

OK Cancel << Hide Help Tool Help



Raster grid of the interpolated surface



631x 467

Developing Datasets without Spatial Analysis

- Gridded datasets can be developed for creating dataset for hydrological simulation models.
- The process requires spatial interpolation using point measurements or observations as a first step.
- When spatial interpolation is executed, data can be estimated at each grid point of interest.

Important steps

- Selection of spatial interpolation method for a particular process variable.
- Visual evaluation of predicted and estimated values need to be evaluated.
- Spatial analysis software such as ArcGIS provides:
 - A visual assessment of predicted and estimated values.
 - Performance measures (errors, etc.)
 - A mechanism to optimize the power of inverse distance weighting method, including the ability to select the nearest neighbors.
 - In case of trend surface models, selection of a specific trend surface model is possible.
 - For Kriging different models are available, with selection of a particular variogram from a number of available authorized variograms.

-
- Few Figures are adopted from copyrighted material from text books from Wiley, TAMU, McGraw Hill and others. Few figures from WWW. Do not distribute or use for commercial purposes.

Spatial Analysis for Water Resources Modeling and Management

Methods for Analysis, Interpretation and Visualization of Spatial Data

Ramesh Teegavarapu, Ph.D., P.E.

Associate Professor,

Director, Hydrosystems Research Laboratory (HRL)

<http://hrl.fau.edu>

Department of Civil, Environmental and Geomatics Engineering,
Florida Atlantic University, Boca Raton, Florida, 33431, USA

Permission to use.

- The material in the presentation is obtained from several copyright protected sources (including journal publications, books and published articles, technical presentations by author and his co-authors). Permission to use the material in this presentation elsewhere needs to be obtained from the author(s) of this presentation as well as publishing agencies which own the copyright permissions for the figures and illustrations.
- Material in the presentation can only be for Academic Use only.
- Journal articles for personal use can be obtained from author: rteegava@fau.edu
- Some of figures are not yet published in any article by the author.
- Any additional information please contact the author :
rteegava@fau.edu

Deterministic and Stochastic Spatial Interpolation

- Topics
 - Basics of spatial interpolation.
 - Point patterns,
 - Spatial autocorrelation,
 - Spatial statistics,
 - Voronoi polygons.
 - Deterministic interpolation methods.
 - Stochastic interpolation methods.
 - Geostatistics, ordinary kriging,
 - Co-kriging and other variants of kriging,
 - Optimal spatial interpolation and spatial interpolation for general of surfaces of hydroclimatic variable surfaces,
 - Missing data estimation, and uncertainty in spatial analysis estimates. Thin plate splines (with or without tension),
 - Locally adaptive interpolation,
 - Trend surface and local polynomial models.
 - Objective selection of points in space for interpolation using optimization methods.

Missing Data Methods (for single site)

- This lecture also provides discussion about a number of new methods developed in the recent years for improved estimation of missing precipitation data at a site.
- Explains simple weighting methods and also stochastic interpolation methods.
- Also introduces RAIN software that is available in public domain.

“ There are **known knowns**; there are things we know we know.
We also know there are **known unknowns**; that is to say we know there are
some things we do not know. But there are also **unknown unknowns** – there
are things we do not know we don't know.”

—United States Secretary of Defense Donald Rumsfeld
Statement during Iraq war

Known Knowns : Observed Data
Known Unknowns: Missing Data
Unknown Unknowns: Mechanisms of missing, etc.

-
- Interpolation
 - Infilling
 - Reconstruction
 - Imputation
 - Gap Filling
 - Filling
 - Prediction
 - Forecasting

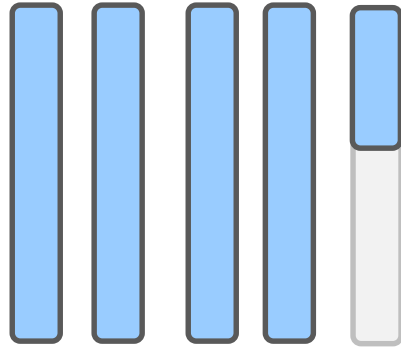
-
- Missing data exist in any hydrological time series
 - Estimation of missing precipitation data remains a critical task that needs to be completed before any hydrologic analysis with serially-complete continuous data is undertaken.
 - Temporal and spatial interpolation methods are often used in infilling or imputation of missing precipitation data.
 - Temporal interpolation using any auto-regressive technique is possible only when serial autocorrelation is high enough for several temporal lags.
 - Climate variability studies require serially complete (gap free) precipitation time series

Missing Data Mechanisms

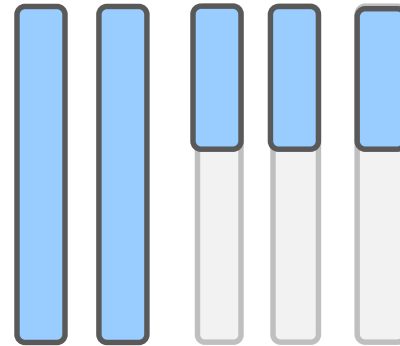
- MAR (**M**issing **A**t **R**andom)
 - Data for a given variable (e.g., Y) are said to be MAR if the **probability** of missing data on Y is unrelated to the value of Y, after accounting for other variables (X).
- MCAR (**M**issing **C**ompletely **A**t **R**andom)
 - Data on Y are said to be MCAR if the **probability** of missing data on Y is unrelated to the value of Y or any values of other variables (X) in a data set.
- MNAR (**M**issing **N**ot **A**t **R**andom)
 - Data on Y are said to MNAR if the **probability** of missing data on Y is related to value of Y or any values of other variables in a data set

-
- In many missing precipitation data estimation studies, gaps can be attributed to be as data missing completely at random (**MCAR**).
 - The number and temporal occurrence of gaps (missing) in precipitation data a site (i.e., rain gauge) are not dependent on the data at the site or any other sites.

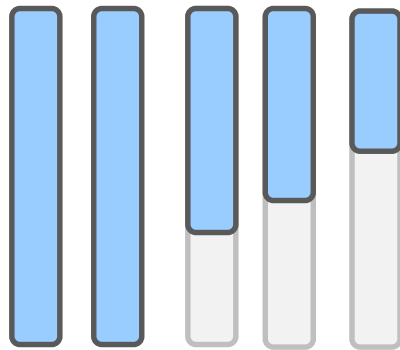
Patterns of missing data



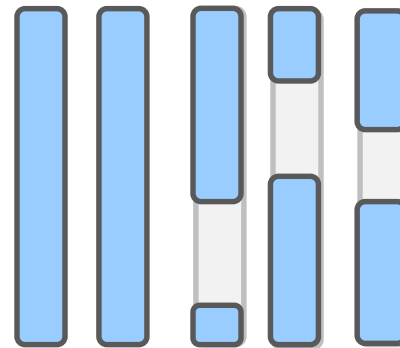
Univariate



Multivariate



Monotone

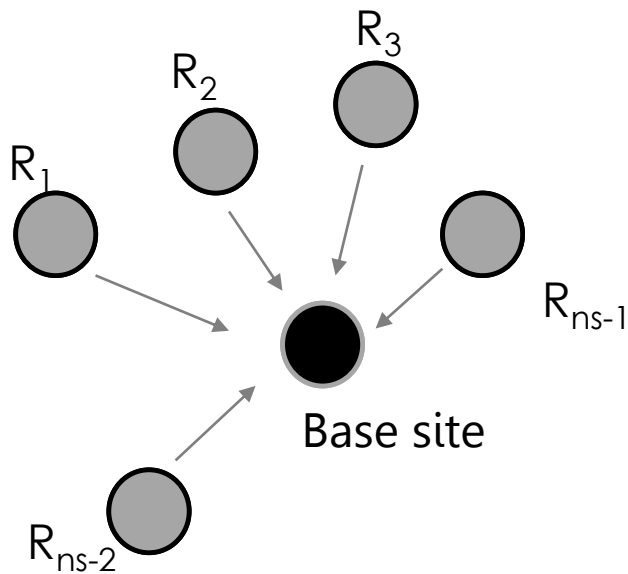


General

-
- Spatial interpolation can be global or local.
 - Spatial interpolation can be exact or inexact.
 - Spatial interpolation can be deterministic or stochastic.

Exact interpolator predicts a value at a known (point location) that is the same as its known value. A surface generated by the interpolator passes through control points.

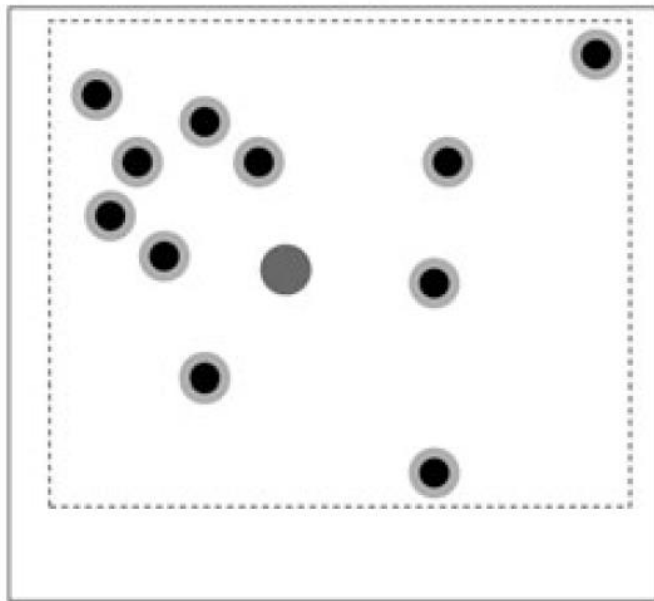
Inexact interpolator predicts a value at a known (point location) that is not the same as its known value. A surface generated by the interpolator may not pass through all control points.



Estimation of missing data at
a single site (base site)
using available data all other
observation sites

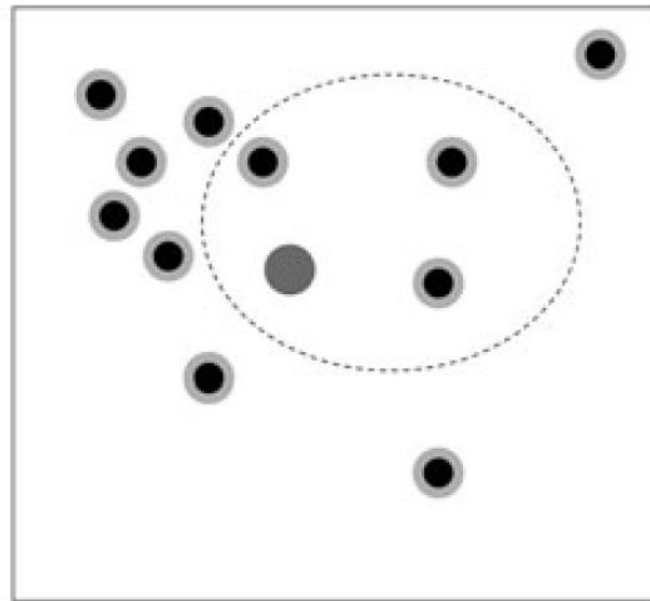
$R_1, R_2, \dots, R_{ns-1}$: Sites with
observations [excluding base site]
Base site: A site with missing data exists
 nf : Number of sites selected.

If $nf = ns-1$ then it is GLOBAL
If $nf < ns-1$ then it is LOCAL



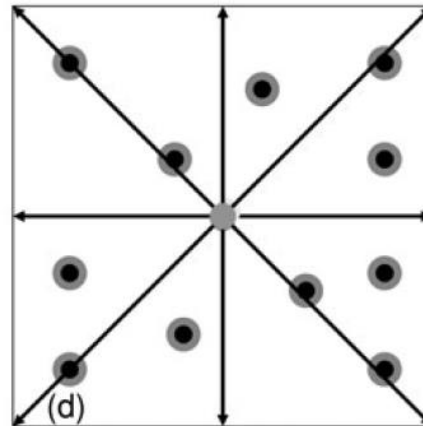
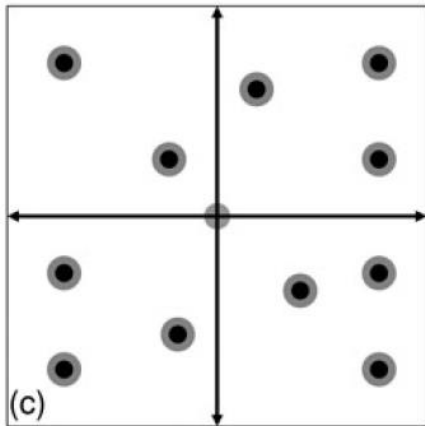
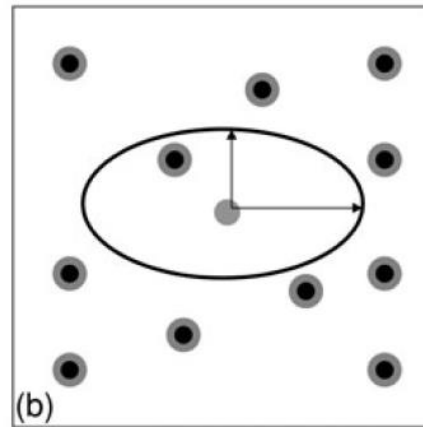
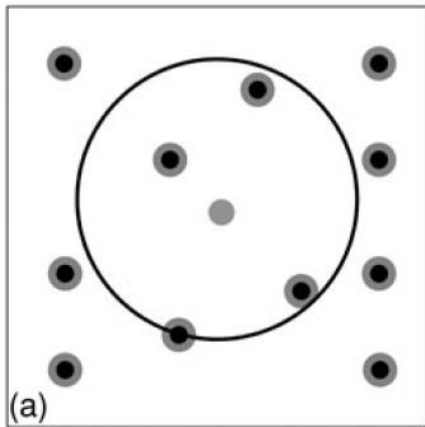
GLOBAL

● Un-sampled Location



LOCAL

● Sampled Location

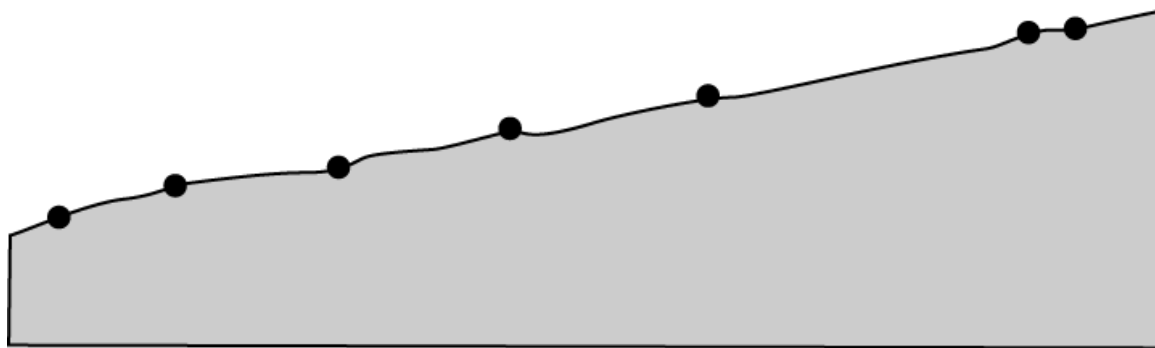


Division of observation space:
Different equal closed polygons to define circular, elliptical, quadrant, and triangular neighborhoods

-
- **Control Points** are points with known values. They provide the data necessary for the development of an interpolator for spatial interpolation.
 - The number and distribution of control points can greatly influence the accuracy of spatial interpolation.

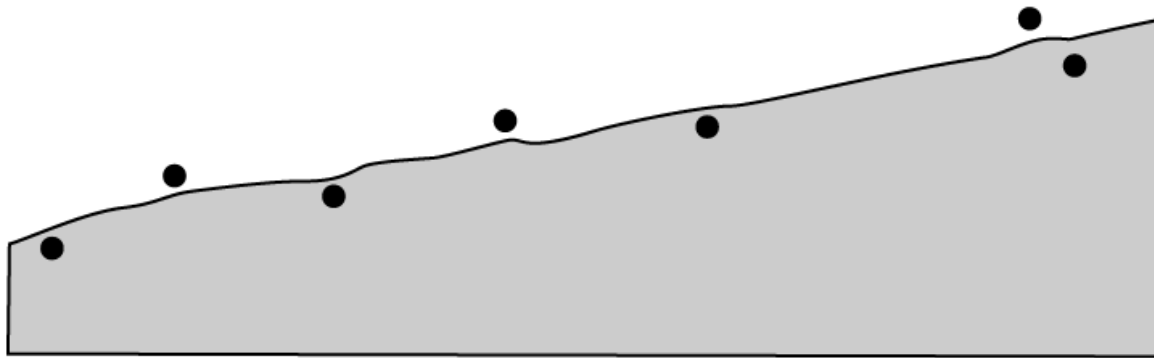
Global		Local	
Deterministic	Stochastic	Deterministic	Stochastic
Trend surface (inexact)*	Regression (inexact)	Thiessen (exact) Density estimation (inexact) Inverse distance weighted (exact) Splines (exact)	Kriging (exact)

Given some required assumptions, trend surface analysis can be treated as a special case of regression analysis and thus a stochastic method (Griffith and Amrhein 1991)



(a)

Exact



(b)

Inexact

Questions:

When would you use Spatial Interpolation ?

When would you use Temporal Interpolation ?

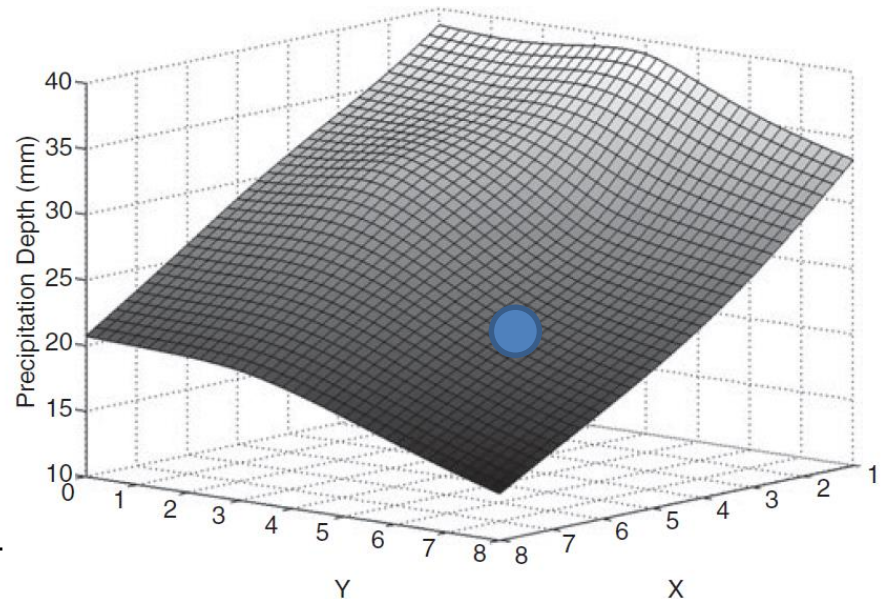
- Temporal Interpolation relies on the **persistence** (via strong **serial autocorrelation**) in the observed time series of variable under consideration.
- Spatial Interpolation depends on **reliable time consistent observations** at different **observation points in space**.
- Spatial Interpolation can also be used for surface generation. Surface generation although aimed for a different purpose can be used for estimating missing data at a point.

-
- Chronological pairing (CP) of data is needed to estimate missing values using spatial interpolation
 - At **low temporal resolutions** observations hydroclimatic variables often show strong persistence allowing for infilling a missing observation in a time interval using observation from the immediate previous time interval (referred to as temporal interpolation).
 - This way of infilling the data is referred to also as **last observation carry forward (LOCF) procedure**. A high serial autocorrelation value at lag one is essential for success of this method.
 - If several missing data exist in a series of continuous time intervals, a constant or a mean value can be replaced for each of the values in these time intervals. This process is referred as the **baseline observation carry forward (BOCF)** approach. BOCF method is also referred to as single imputation.

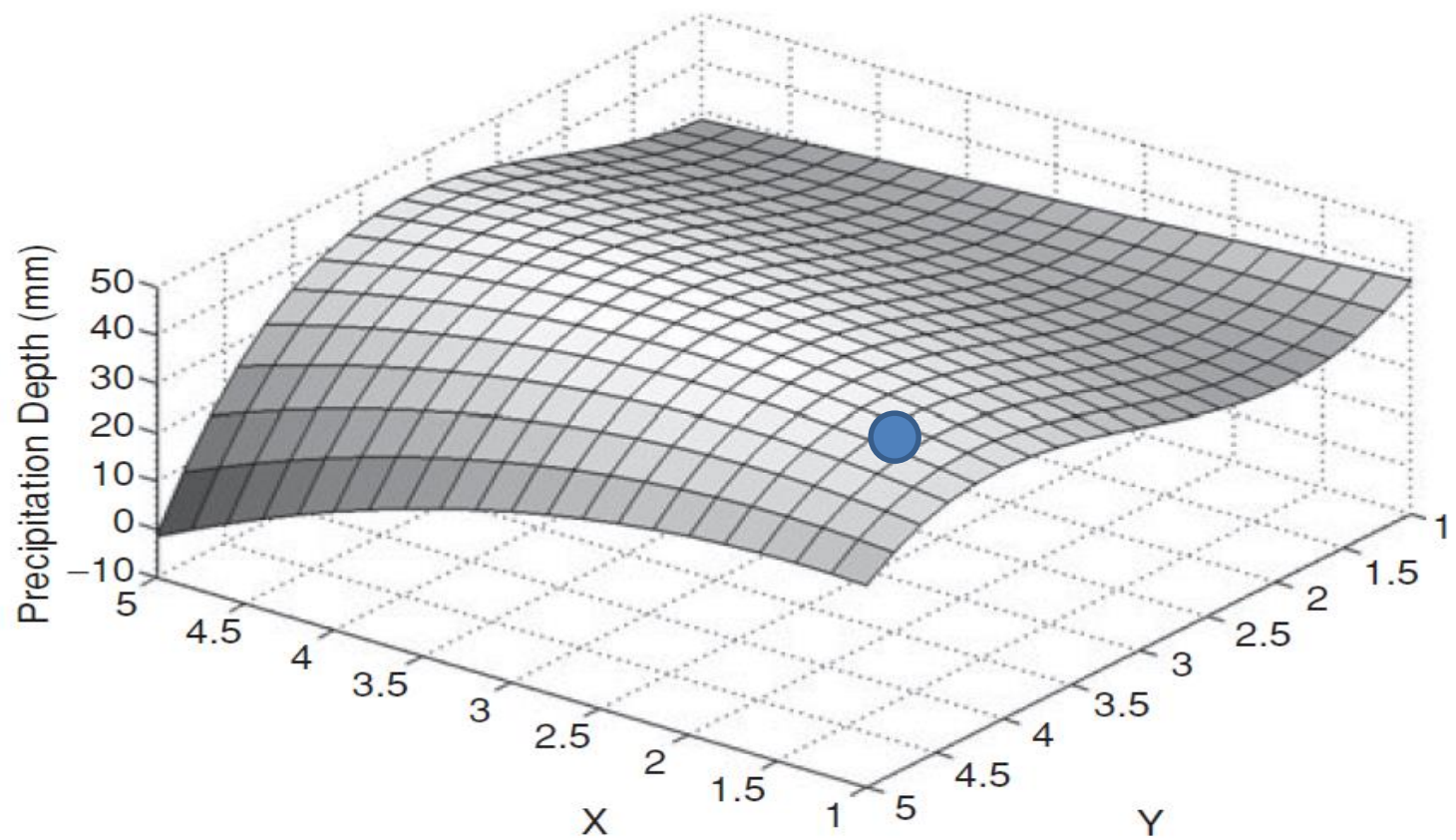
-
- Focus of this talk will be on “ [Spatial Interpolation Methods](#)” and several conceptual variants developed for improved estimation of precipitation data.
 - The temporal scale (e.g. day or larger) at which data are filled requires spatial interpolation as opposed to temporal interpolation.
 - Reasons: Low autocorrelation at multiple lags (including lag-1 correlation) – Lack of persistence
 - Gaps may not be part of one storm (or a single event)
 - Spatial interpolation serves well when several observation sites are located nearby and have reliable observations including gap-free and error-free.

-
- 1) What is the influence of spatio-temporal precipitation patterns on estimation methods?
 - (2) How do we quantify and delineate the observation space (number of points, clusters, etc.)?
 - (3) How do we capture the correlation structure of the observations from a monitoring network for estimation of missing data?
 - (4) How do we use auxiliary data (other than traditional rain gage observations) to improve the estimation?

-
- Inverse Distance Weighting Method (also NWS method)
 - Normal Ratio Method
 - Quadrant Method
 - Different forms of Kriging
 - Trend surface Models
 - Local and Global
 - Thin Plate Splines



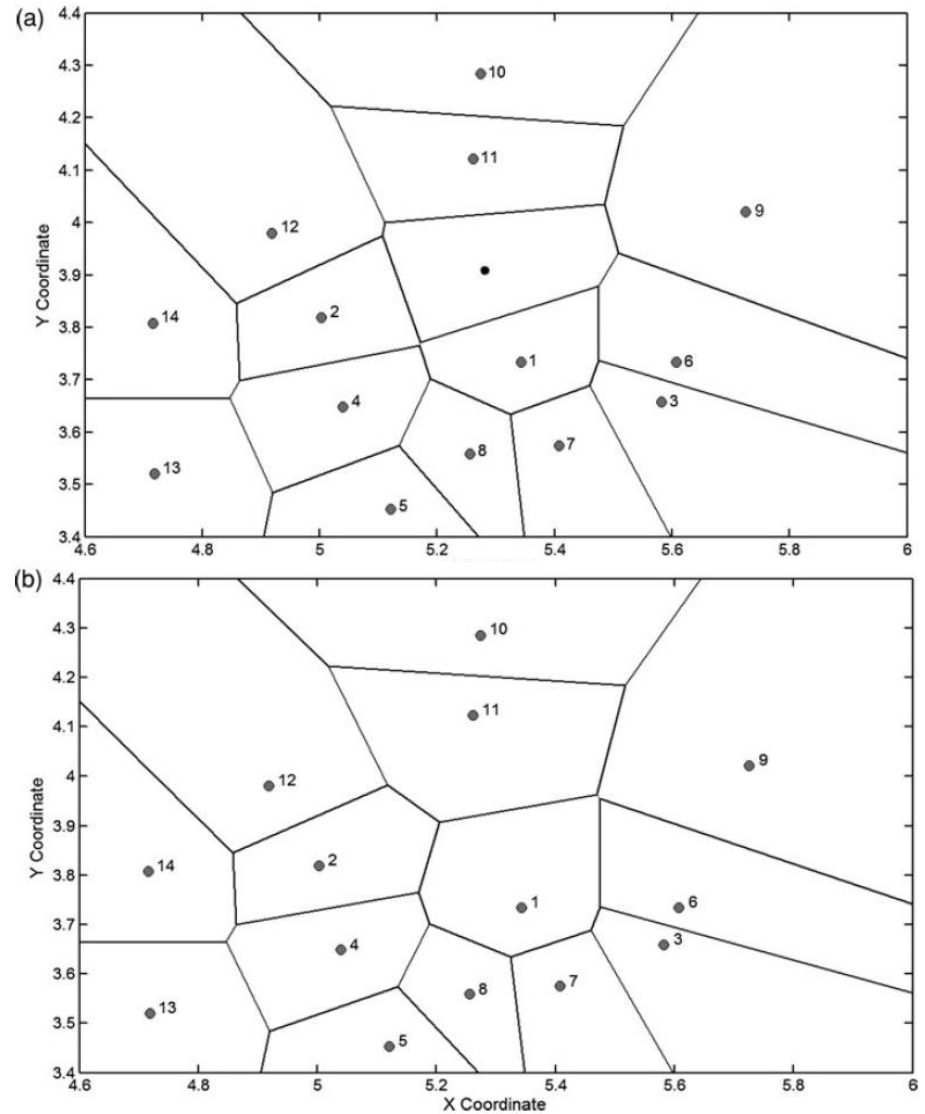
● Point of Interest

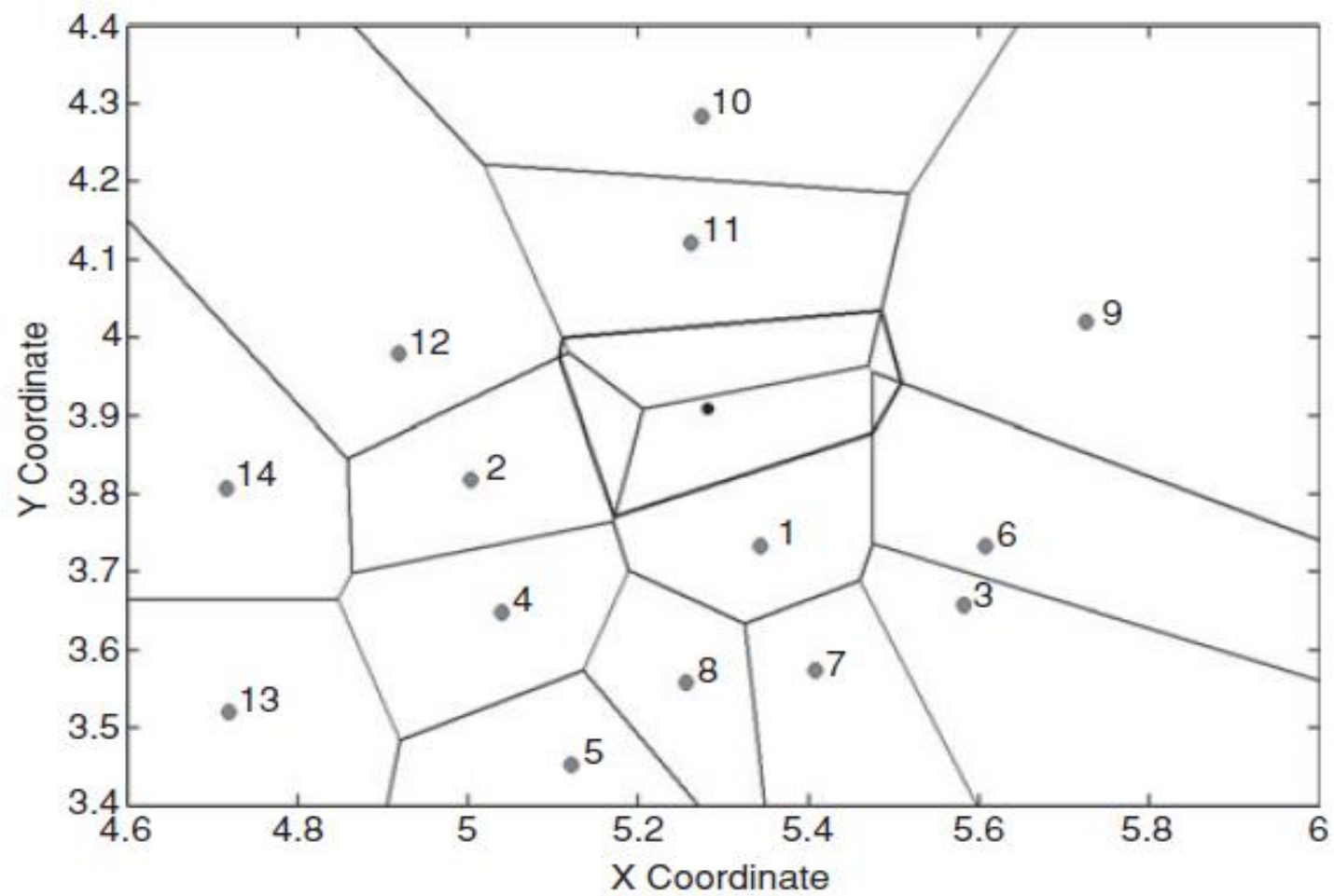


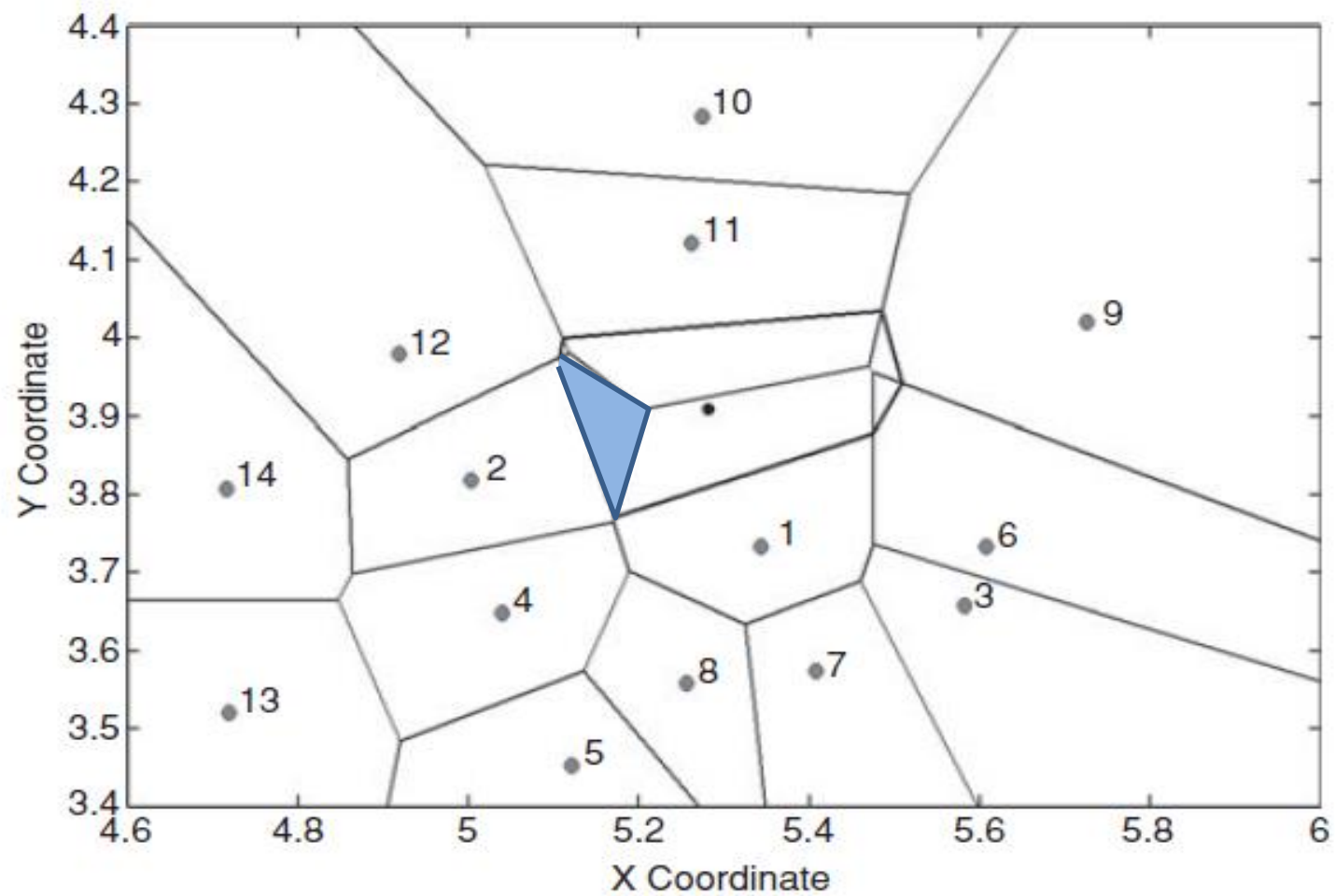
$$w_{m,j} = \frac{A_m}{A_j} \quad \forall j$$

$$\theta'_{m,n} = \sum_{j=1}^{ns-1} w_{m,j} \theta_{j,n} \quad \forall n$$

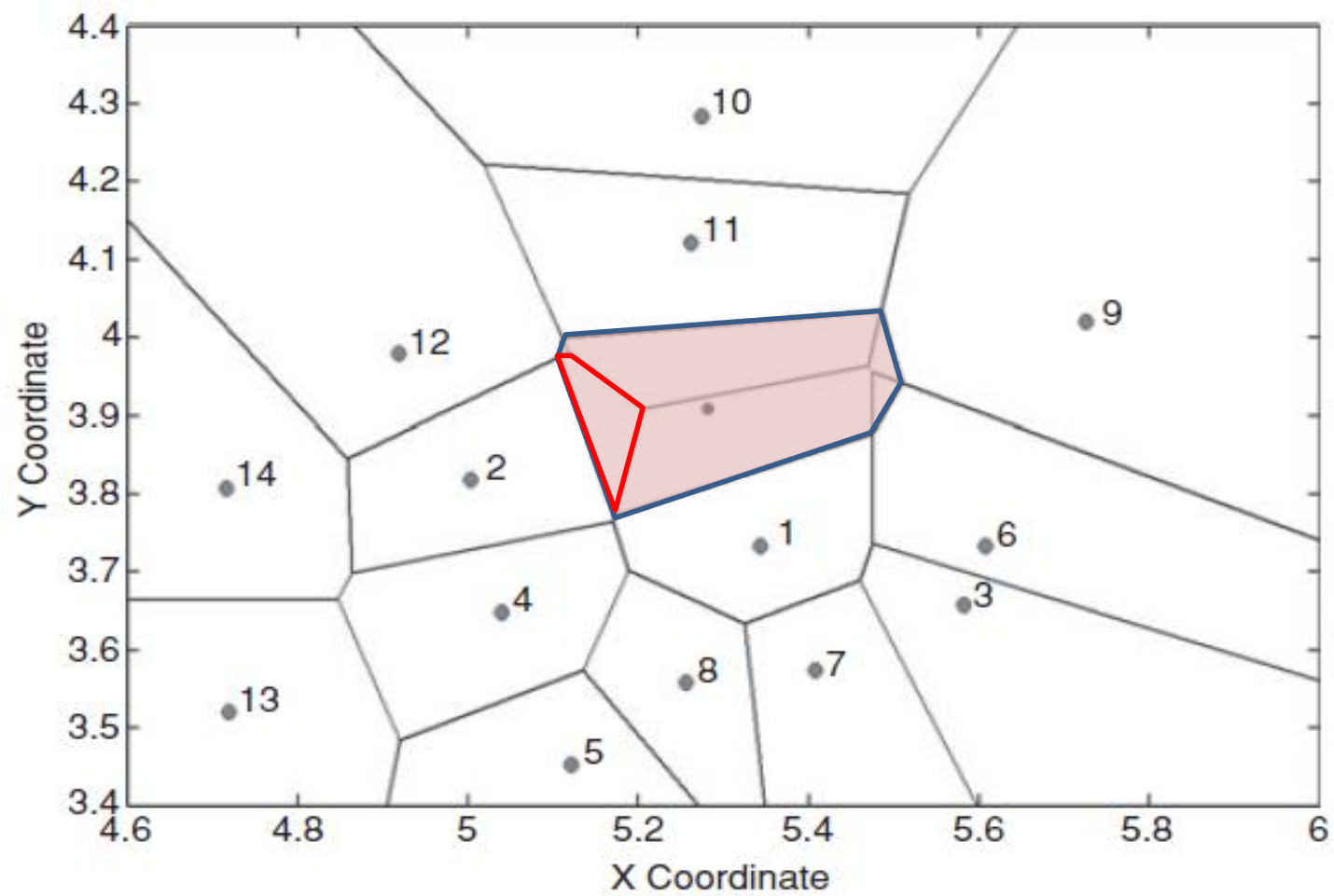
Thiessen Polygon
without the site
with missing data



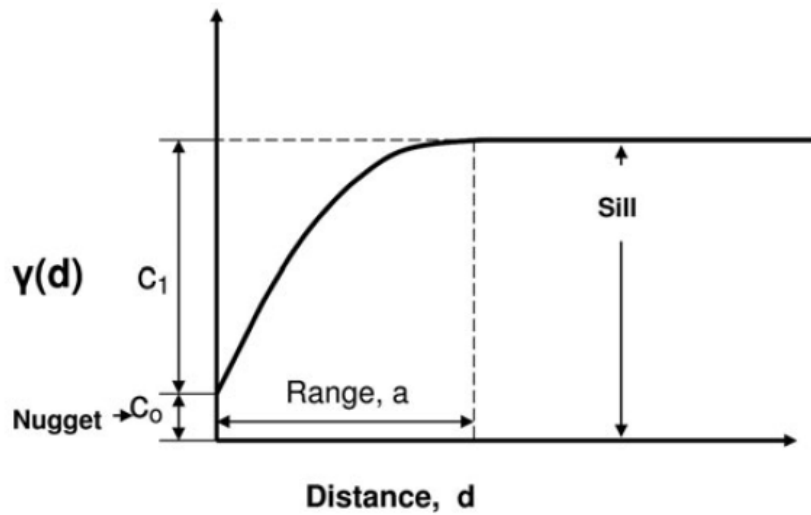




A_2

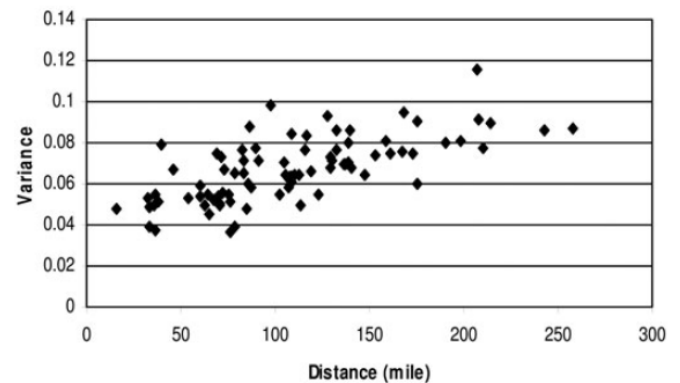


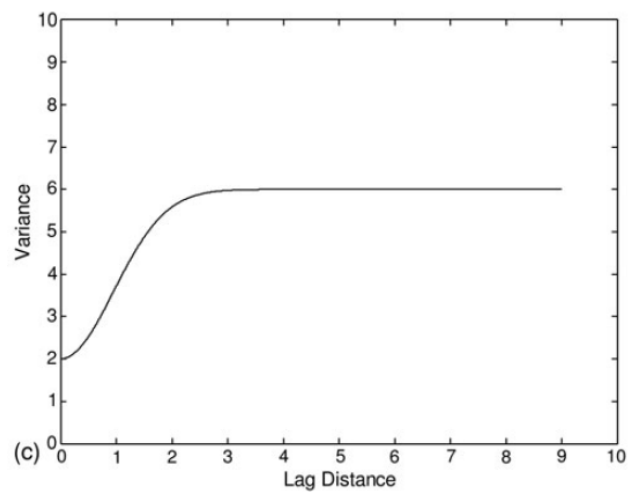
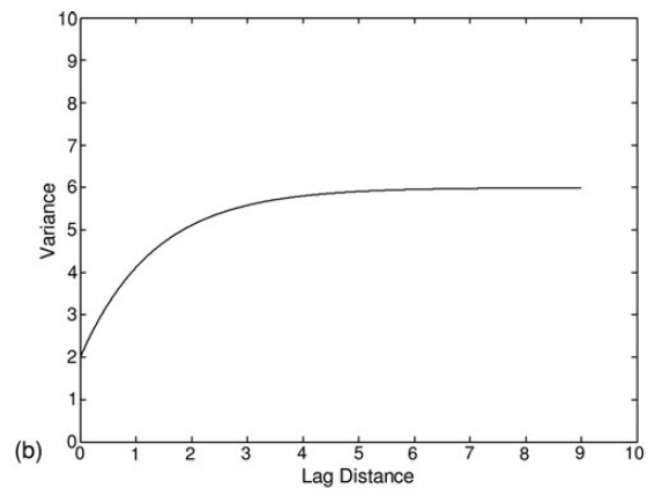
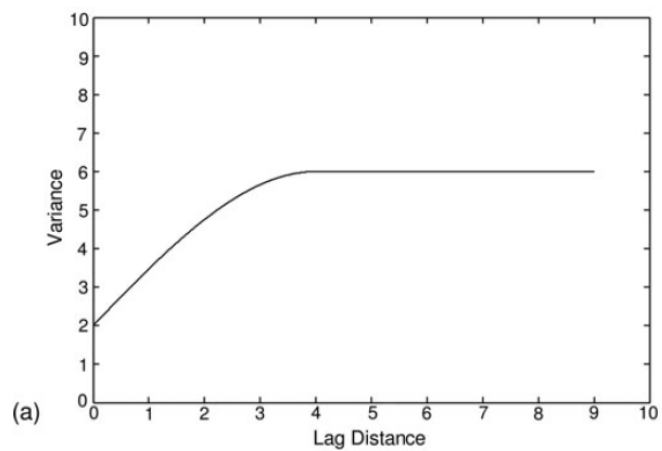
A_j



$$\gamma(d) = \frac{1}{2n(d)} \sum_{d_{ij}} (\theta_i - \theta_j)^2$$

An example
Variogram cloud





$$\gamma(d)_1 = C_o + C_1 \left[\frac{1.5d}{a} - 0.5 \left(\frac{d}{a} \right)^3 \right]$$

$$\gamma(d)_2 = C_o + C_1 \left[1 - \exp \left(-\frac{3d}{a} \right) \right]$$

$$\gamma(d)_3 = C_o + C_1 \left[1 - \exp \left(-\frac{(3d)^2}{a^2} \right) \right]$$

$$\gamma(d)_4 = C_o + C_1 \left[1 - \frac{2}{\pi} \cos^{-1} \left(\frac{d}{a} \right) + \frac{2d}{\pi a} \sqrt{1 - \frac{d^2}{a^2}} \right]$$

Spherical

Exponential

Gaussian

Circular

-
- Variants developed include:
 - Modified Inverse Distance Method
 - Integrated Distance and Thiessen Polygon Weighting Method
 - Correlation Weighting Method
 - Nearest Neighbor (based on patterns) Method
 - Universal Approximate Ordinary Kriging Method (UOK)
 - Positive Kriging Method
 - Artificial Neural Network Method
 - Optimal Spatial Weighting Method
 - Optimal Spatial Weighting Method with ability to select optimal set of neighbors.
 - Optimal Set of neighbors –Inverse Distance Method, Normal Ratio Method and Quadrant Method

-
- Variants developed include:
 - Cluster-based Optimal Weighting Method
 - Proximity-Metric based Optimal Weighting Method
 - Optimal Classification-based Method
 - Site and Regional Statistics Preserving Method
 - Optimal Single Best Estimator (SBE) Method
 - Optimal Weighting Method with SBE modifications
 - Multiple Imputation Method
 - Optimal Weighting Method integrated with Statistical Corrections
 - Areal Weighting methods for Grid-based Transformations of Precipitation Estimates

-
- Most of the techniques discussed directly relate to **point observations alone**. Therefore, they cannot be used for surface generation.
 - Substantial length of data (historical data) is available for estimation of parameters in the methods.
 - All the formulations require constant spatial arrangement of gages over time and observations are available at all other locations used in spatial interpolation.
 - If not, **Filler algorithms** are required before spatial interpolation can be carried out.
 - Uncertainty in estimates needs to be assessed.
 - The techniques need to be tested at finer time resolutions
 - Assessment of outputs from calibrated hydrologic simulation models using estimates of missing precipitation data from different approaches is being carried out.

-
- Procedures for **objective** selection of **optimal** number of neighbors (or gauges) and **neighborhood size** (in reference to local or global interpolation);
 - **Optimal weights** in distance or correlation-based or other weighting schemes
 - The ability to **preserve site-specific statistics** of precipitation data;
 - Ability to provide uncertainty in the estimates,
 - Mechanism to **define rain or no-rain conditions** at the site of interest or correct the estimates based on these conditions;
 - Ability to preserve **all the precipitation characteristics** (e.g., dry and wet spells, extremes, transitions (e.g., wet to dry), autocorrelation and several others)
 - Ability to preserve **spatial variance** and **regional statistics**.

“The fact that conceptually simple spatial interpolation and weighting methods some times perform better than complex stochastic methods should not deter research in developing **new conceptually appealing and sound methods**.

These new methods besides offering innovative approaches use traditional and **emerging computational paradigms** to understand the structure of the spatial data. Capturing this **spatial structure** of data by mechanisms other than **distances** and **variances** needs to be explored. Knowing the unknowable should not be limited to known methods”

Missing Precipitation Estimation and Related Works

- **Ramesh S. V. Teegavarapu** and V. Chandramouli, Improved Weighting Methods, Deterministic and Stochastic Data-driven Models for Estimation of Missing Precipitation Records, Journal of Hydrology, 191-206, 312, 2005 [Link to Publication](#)
- **Ramesh S. V. Teegavarapu**, Spatial Interpolation using Non-linear Mathematical Programming Models for Estimation of Missing Precipitation Records, Hydrological Sciences Journal, 57(3), 383-406, 2012. [Link to Publication](#)
- **Ramesh S. V. Teegavarapu**, Statistical Corrections of Spatially Interpolated Precipitation Estimates, Hydrological processes, 28(11), 3789–3808, 2014. DOI: 10.1002/hyp.9906. [Link to Publication](#)
- **Ramesh S. V. Teegavarapu**, Missing Precipitation Data Estimation using Optimal Proximity Metric-based Imputation, Nearest Neighbor Classification and Cluster-based Interpolation Methods, Hydrological Sciences Journal, 2013. DOI:10.1080/02626667.2013.862334. [Link to Publication](#)
- **Ramesh S. V. Teegavarapu**, Estimation of Missing Precipitation Records Integrating Surface Interpolation Techniques and Spatio-Temporal Association Rules , Journal of HydroInformatics Vol, 11 No 2 pp 133–146, 2009. [Link to Publication](#)
- **Ramesh S. V. Teegavarapu**, Tadesse Meskele and Chandra Pathak, Geo-Spatial Grid-based Transformation of Multi-Sensor Precipitation using Spatial Interpolation Methods, Computers and Geosciences, 40, 28-39, 2012. Doi:10.1016/j.cageo.2011.07.004. [Link to Publication](#)
- **Ramesh S. V. Teegavarapu**, Use of Universal Function Approximation in Variance-dependent Interpolation Technique: An Application in Hydrology, 332, 16-29, 2007. [Link to Publication](#)
- **Ramesh S. V. Teegavarapu**, Mohammad Tufail and Lindell Ormsbee, Optimal Functional Forms for Estimation of Missing Precipitation Records, Journal of Hydrology, 2009, 374 (2009) 106–115. [Link to Publication](#)

Works

- **Ramesh S. V. Teegavarapu**, Aneesh Goly, Qinglong Wu, A Comprehensive Framework for Assessment of Radar-based Precipitation Data Estimates, Journal of Hydrologic Engineering, ASCE, 2015. [Link to Publication](#)

Books

- **Ramesh S. V. Teegavarapu**, Floods in a Changing Climate: Extreme Precipitation, , Cambridge University Press-UNESCO (United Nations Educational Scientific and Cultural Organization), January, 2013. 285 pages.
- Manual of Standard Practice for Radar Rainfall Data Estimation, Editors: Chandra Pathak and **Ramesh S. V. Teegavarapu**, completed August 2016, ASCE. Available:

Book Chapters

- **Ramesh S. V. Teegavarapu**, Spatial and Temporal Estimation and Analysis of Precipitation, Handbook of Applied Hydrology, McGraw Hill, in Print/publication.
- **Ramesh S. V. Teegavarapu**, Precipitation, Chapter in ASCE Book on Statistical Distributions in Hydrology: Ramesh Teegavarapu and Chandra Pathak, September 2016, ASCE. Under Publication.

Evaluation of Interpolated Estimates

The spatial interpolation methods can be evaluated using error measures and performance measures such as:

- Root mean squared error (RMSE),
- Mean absolute error (MAE)
- Goodness-of-fit measure criterion, coefficient of correlation (ρ) or determination (R^2), based on observed and estimated values.
- Factors such as intuitive reasonableness of model or approach, conceptual accuracy and simplicity of the model are considered to make an objective assessment, and finally selection of the models or approaches used for estimation of missing data

-
- **User-specified weights can be attached to performance metrics or measures as numerical values or *fuzzy membership function-based values*. (to quantify the importance of one measure over the other. In addition to these performance measures,**
 - **A bias plot or analysis of bias is also recommended as a means of understanding the structure of errors,**
 - **Model specific bias and distribution of errors over a specific range of values or time.**

Other Evaluation Measures

- A number of performance measures, error measures and statistical hypothesis tests to evaluate different spatial interpolation methods.
- Summary statistics, quantile-quantile (Q-Q) plots, statistical hypothesis tests (e.g., two sample Kolmogorov-Smirnov, Chi-square tests) to check the similarity of probability distributions,
- Differences in two state first-order Markov chain transition probabilities for dry-wet, dry-dry, wet-dry and wet-wet conditions,
- Autocorrelations at several temporal lags and deviations in extreme precipitation indices derived based on observed and estimated datasets.
- Extreme Precipitation Indices

Homogeneity Tests

- **Homogeneity of the filled data sets should be tested using one of the tests.**
- **Pettitt's (Pettitt, 1979),**
- **Buishand's (Buishand, 1982)**
- **Alexandersson's standard normal homogeneity test (SNHT) (Alexandersson, 1986)**
- **Von Neumann ratio test (Von Neumann, 1941).**

Checks

- Evaluation of residuals based on observed and estimated values
- by interpolation methods is essential.
- Independence of residuals,
- Normality of residuals,
- Homoscedasticity.
- Independence of residuals can be assessed by autocorrelation plots.
- Normality of residuals is checked by evaluating the histogram of the residuals and probability plots.
- Homoscedasticity (i.e., constant error variance) or heteroscedasticity (i.e., non-constant error variance) can be assessed by plotting standardized residuals and estimated values.
- A check for heteroscedasticity of residuals is essential and can be done using different tests: e.g. Breush–Pagan test or the White test.

-
- Random and Systematic Errors
 - Spatial interpolation Algorithms – Deterministic and Stochastic Models are generally used for estimation of missing precipitation data
 - Each method has several advantages and limitations
 - IDWM – tent pole effect
 - Kriging – Selection of Variogram, isotrophic effects
 - Thin Splines
 - Local Polynomial functions
 - Regression Models
 - ANN
 - Optimal Functional forms

-
- Local or Global Control points for Interpolation
 - Existence of Clusters
 - Rain or No Rain pre-predictions
 - Availability of historical Data
 - Spatial Variability of Rainfall
 - Availability of other sources of precipitation data

Tobler's first law of geography (1970) :

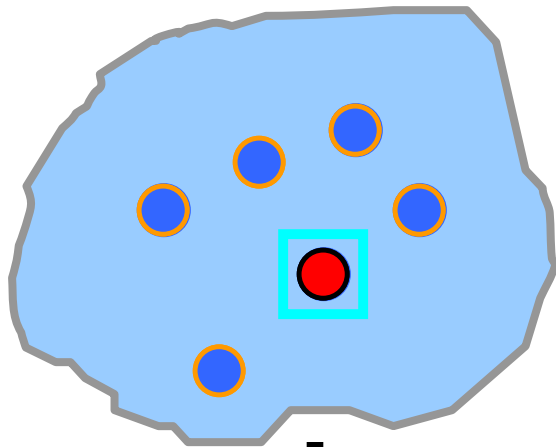
*“Everything is related to everything else,
but near things are more related than
distant things”,*

which forms the basis for many interpolation techniques.

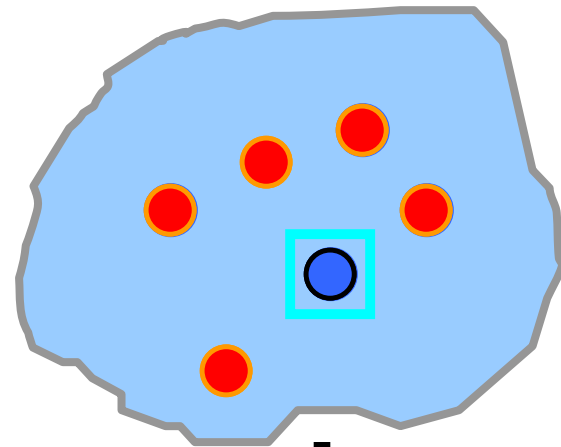
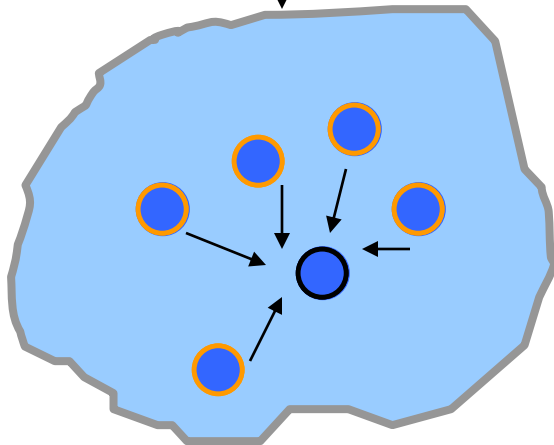
-
- All interpolation techniques fail in estimation of missing precipitation data at a point in space in two situations:
 - 1) when precipitation is measured at all or few other stations and no precipitation occurred in reality at the base station; and
 - 2) when precipitation is measured at the base station and no precipitation is measured or occurred at all the other stations.
 - **Base station:** Station with missing precipitation records

-
- In case 1, all spatial interpolation techniques provide a **positive** value of estimate while in reality a **zero** value of precipitation is recorded at the base station.
 - It is impossible to estimate missing precipitation data in the second case as the point observations are used to estimate the missing value at the base station by using spatial interpolation algorithms alone. All the interpolation techniques provide a **zero value** as an estimate for situations encountered in case 2

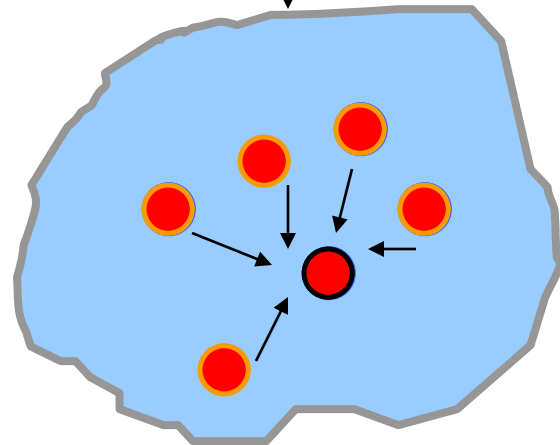
Case I



Interpolation

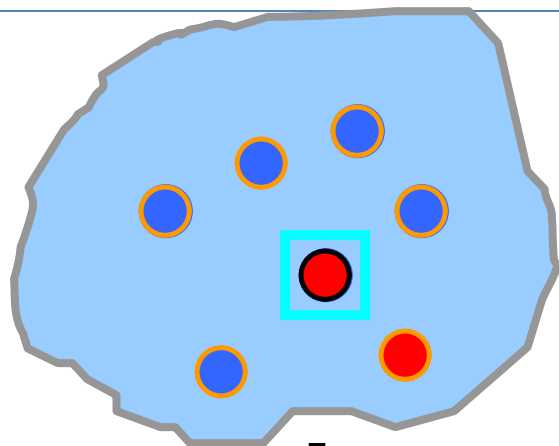


Interpolation

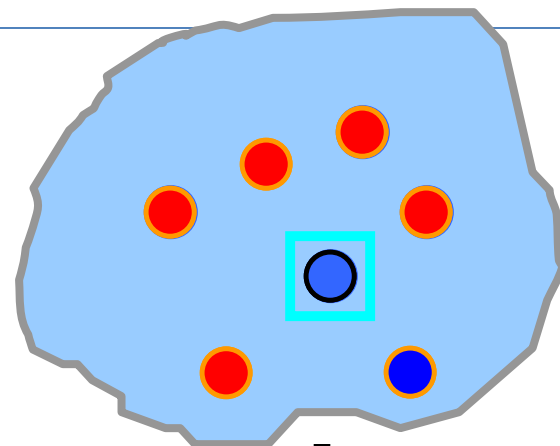
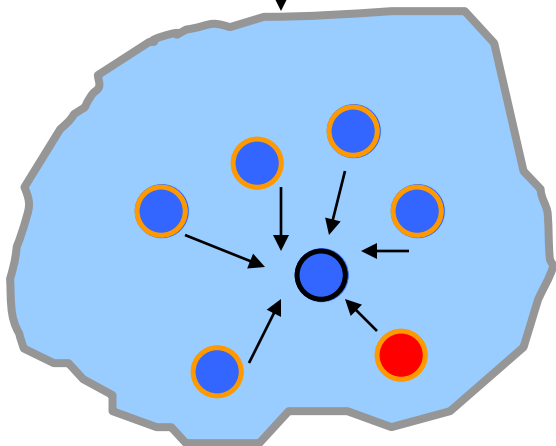


● Positive value

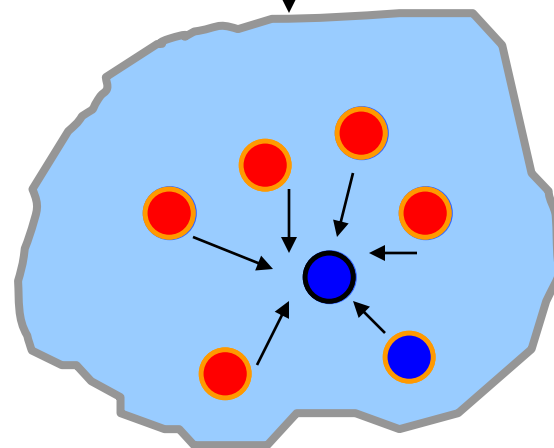
● Zero



Interpolation



Interpolation



● Positive value

● Zero

-
- Single Best Estimator (SBE)
 - Distance, correlation
 - Observed value – nearest to gage (Euclidean distance)
 - Observed value from highest correlated gage.
 - Gage Mean Estimator (GME)
 - Local or Global
 - Average of rain gage observations
 - Climatological Mean Estimator (CME)
 - Average value of rain gage observations for a specific time interval based on historical data

-
- Improvements and Improvisations of traditional weighting methods
 - Correlation based weighting
 - Optimal Correlation based weighting.
 - Use of Artificial Intelligence Techniques
 - Universal Function Approximators (ANNs)
 - Universal functional approximation based Kriging
 - Optimal Functional Forms using Genetic Algorithms
 - Knowledge Discovery Methods
 - Data Mining – Association Rule Mining (ARM)
 - Mathematical Programming Models
 - Optimal Weighting Methods
 - Optimal Spatial cluster based weighting method
 - Optimal post Kriging models

-
- Selection of control points in space for interpolation
 - Local or Global
 - Selection of clusters (not grouped at one location in space)
 - Euclidean distance is not always a surrogate for spatial autocorrelation
 - Isotropic issues
 - Lack of historical spatial and temporal data
 - To characterize the spatial structure of the data

-
- Spatial interpolation is based on understanding the spatial structure of the data
 - Weights are based on
 - $W = f(d)$
 - $W = f(\rho)$
 - $W = f(\text{variance})$
 - Creation of surfaces or fields are not required for the current problem

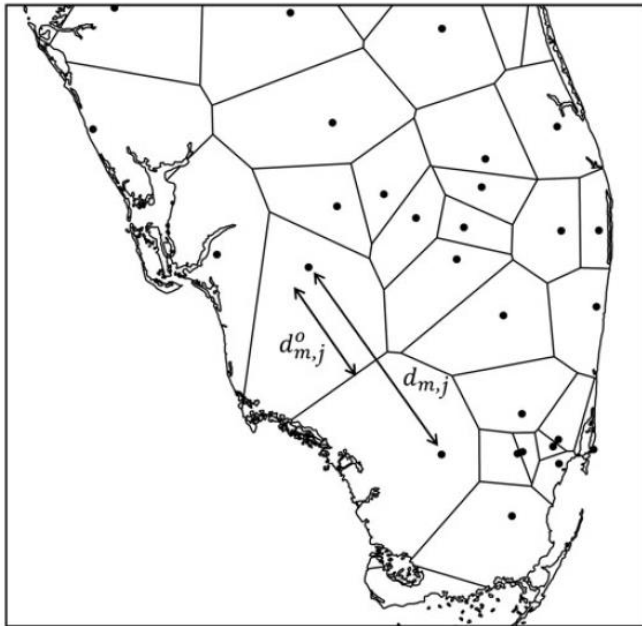
Variants of Traditional Weighting Methods

$$\theta_m = \frac{\sum_{i=1}^n \theta_i \rho_{mi}}{\sum_{i=1}^n \rho_{mi}}$$

$$\theta_m = \frac{\sum_{i=1}^n \theta_i (\rho_{mi})^k}{\sum_{i=1}^n (\rho_{mi})^k}$$

The exponent can be optimized

- Based on the property of Thiessen Polygon, distances for the reciprocal distance method are modified.

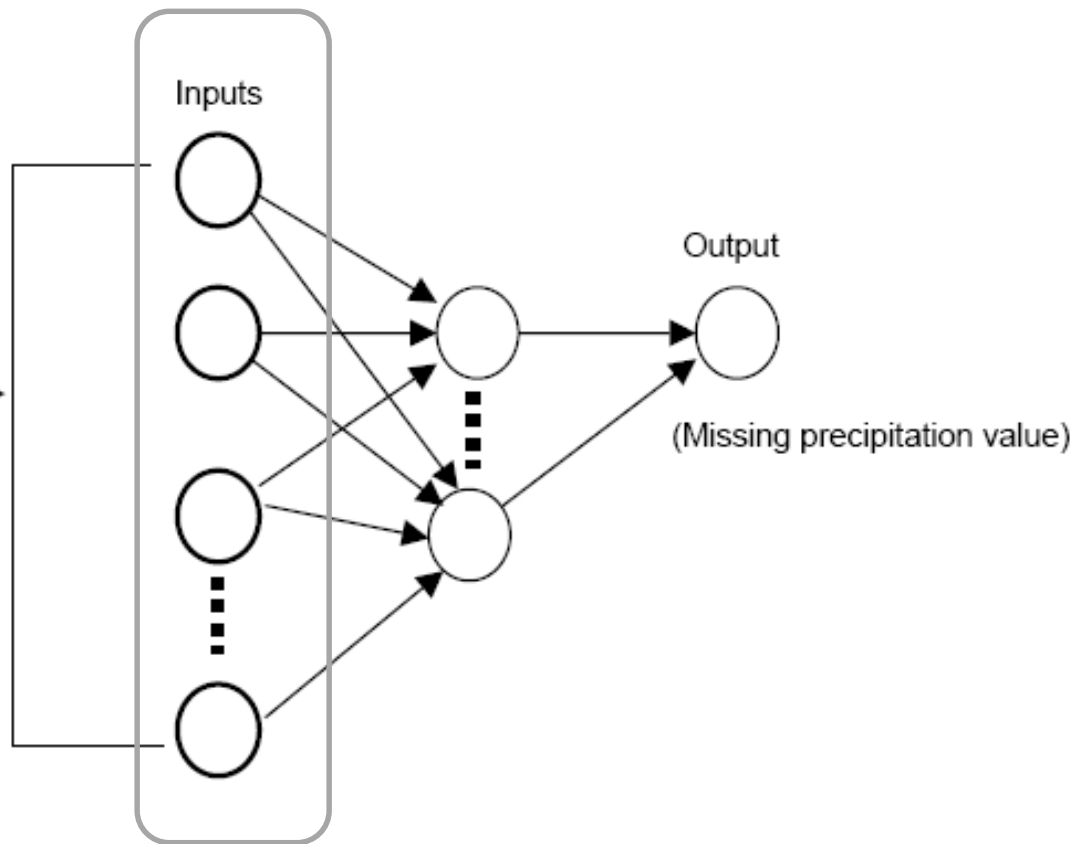


The inverse distance weighting method is modified by replacing the distance by a distance defined by the property of the proximity (Thiessen) polygons.

The distance is smaller than the actual distance measured from the base station.

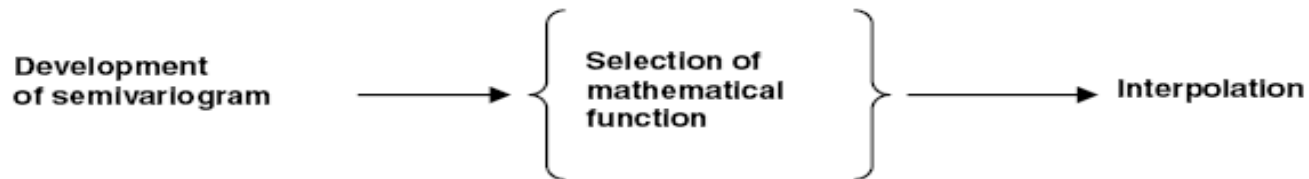
-
- Artificial Neural Networks
 - Inductive Modeling
 - Data-driven model – Need Historical Data
 - Connection Weights - Decide which stations will participate in the function approximation
 - Limitations
 - Negative estimates
 - Need re-adjustment of negative values
 - Very Sensitive to Training data, data length, data representativeness – Station history

Precipitation
values
recorded at
stations

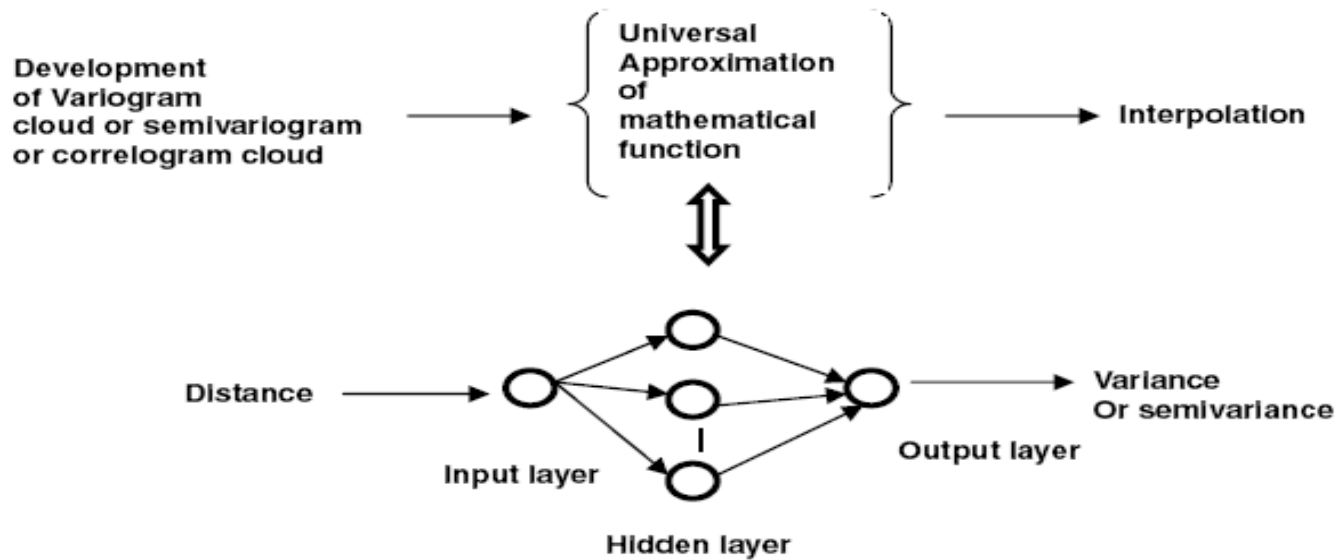


+Highest Correlated Stations

-
- Traditional kriging has a major limitation due to the need for an **a priori definition of a mathematical function for a semivariogram** that might fit the surface to be interpolated.
 - Use of the universal function approximator, artificial neural network (ANN), as a replacement to **fitted authorized semivariogram model** within ordinary kriging was investigated.
 - All the conditions required for authorized semi-variogram should be met before UOK can be applied.



a Traditional ordinary Kriging



b Universal function approximation based Kriging

-
- Post modification of Kriging derived weights.
 - Formulation

Minimize

$$\sum_{j=1}^n \left(\sum_{i=1}^N (\lambda_i \theta_i^j) - \theta_m^j \right)^2$$

Subject to :

$$\sum_{i=1}^N \lambda_i = 1$$

$$\lambda_i \geq 0 \forall i$$

Post optimization scheme is employed to obtain positive weights.

-
- Simple optimization formulations were used by several researchers in spatial interpolation (Gandin, 1965, Ahrens, 2006)
 - Optimal Functional form
 - Derivation of functional forms that relate data at control point in space where data is missing.
 - Optimal Kriging for weights
 - Kriging with optimization

Optimal Functional Forms

Use of Genetic Algorithms to obtain optimal functional forms with

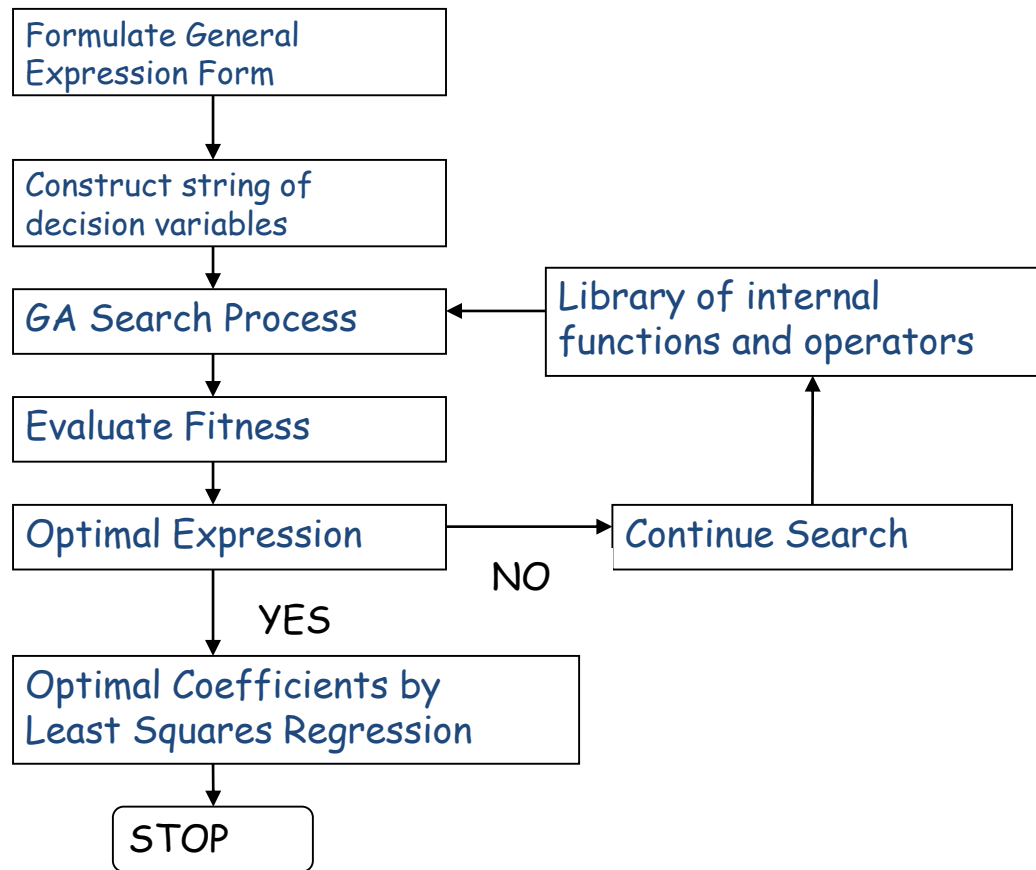
- Mathematical Operators
- Simple functional forms
- Linear/non-linear optimization to obtain the coefficients in the identified Functional forms

$$Y = \textcircled{a_1} \textcircled{F(X_1)} \textcircled{@} \textcircled{b_1} \textcircled{F(X_2)}$$

Coefficients $a_1, b_1 = \{\text{real numbers}\}$

Mathematical Operator $@ = \{+, -, *, /, ^\}$

$F() = 0, 1, X, \log(X), \exp(X), \text{trig}(X), \sin(X), \text{sqrt}(X), 1/X, \text{etc.}$



Operator #	Operator
1	+
2	-
3	*
4	/
5	^

$$\theta_m = \frac{\sum_{i=1}^n \theta_i * (FFSGA \text{ Functional Form})_i}{\sum_{i=1}^n (FFSGA \text{ Functional Form})_i}$$

Functions for Decision Variables: Distance (d_{mi}) and correlation coefficient (R_{mi})

Function #	Function $f(d_{mi})$ or Function $f(R_{mi})$
1	1
2	d_{mi} or R_{mi} or $\text{Sqrt}(d_{mi})$ or $\text{Sqrt}(R_{mi})$
3	$1/d_{mi}$ or $1/R_{mi}$
4	$\text{Exp}(d_{mi})$ or $\text{Exp}(R_{mi})$
5	$\text{Log}_e(d_{mi})$ or $\text{Log}_e(R_{mi})$
6	$\text{Log}_{10}(d_{mi})$ or $\text{Log}_{10}(R_{mi})$
7	$\text{Exp}(1/d_{mi})$ or $\text{Exp}(1/R_{mi})$
8	$\text{Log}_e(1/d_{mi})$ or $\text{Log}_e(1/R_{mi})$
9	$\text{Log}_{10}(1/d_{mi})$ or $\text{Log}_{10}(1/R_{mi})$
10	$d_{mi} * \text{Exp}(d_{mi})$ or $R_{mi} * \text{Exp}(R_{mi})$
11	$d_{mi} * \text{Log}_e(d_{mi})$ or $R_{mi} * \text{Log}_e(R_{mi})$
12	$d_{mi} * \text{Log}_{10}(d_{mi})$ or $R_{mi} * \text{Log}_{10}(R_{mi})$
13	$(1/d_{mi}) * \text{Exp}(d_{mi})$ or $(1/R_{mi}) * \text{Exp}(R_{mi})$
14	$(1/d_{mi}) * \text{Log}_e(d_{mi})$ or $(1/R_{mi}) * \text{Log}_e(R_{mi})$
15	$(1/d_{mi}) * \text{Log}_{10}(d_{mi})$ or $(1/R_{mi}) * \text{Log}_{10}(R_{mi})$

$$\theta_m = \frac{\sum_{i=1}^{14} \theta_i C_i \left[C_1 \left[R_{mi} \log_{10} \left(\frac{1}{R_{mi}} \right) - \left(\frac{1}{R_{mi}} \right) \log_{10} (R_{mi}) \right]^{\wedge} C_2 \left[\frac{\log_{10} \left(\frac{1}{d_{mi}} \right)}{\log_{10} (d_{mi})} \right] \right]}{\sum_{i=1}^{14} C_i \left[C_1 \left[R_{mi} \log_{10} \left(\frac{1}{R_{mi}} \right) - \left(\frac{1}{R_{mi}} \right) \log_{10} (R_{mi}) \right]^{\wedge} C_2 \left[\frac{\log_{10} \left(\frac{1}{d_{mi}} \right)}{\log_{10} (d_{mi})} \right] \right]}$$

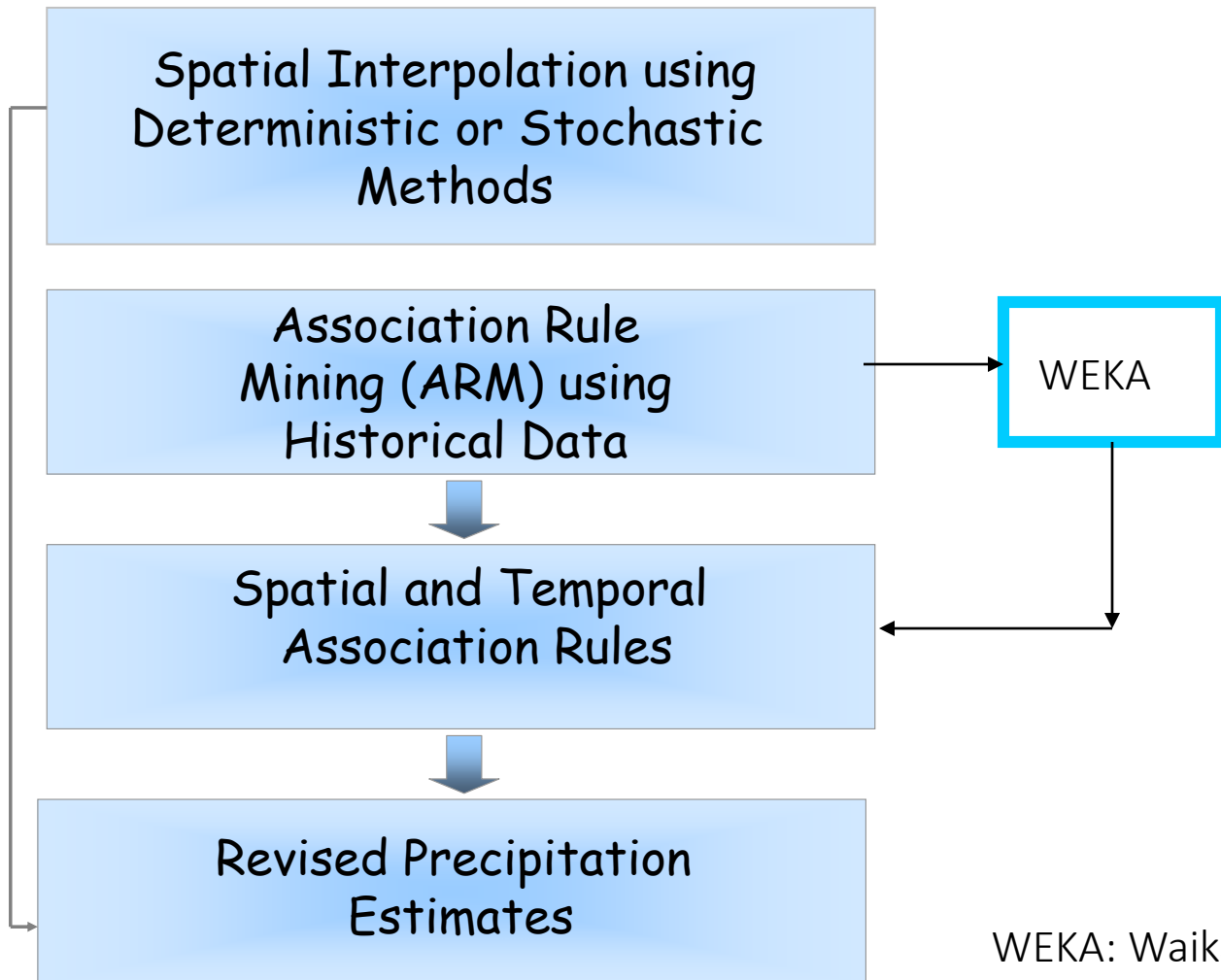
$$\theta_m = \frac{\sum_{i=1}^{14} \theta_i C_i \left[\left[\frac{\exp(R_{mi})}{\left(\frac{1}{R_{mi}} \right) \log_{10} \left(\frac{1}{R_{mi}} \right)} \right] + \left[\sqrt{d_{mi}} \log_{10} \left(\frac{1}{d_{mi}} \right) \right] \right]}{\sum_{i=1}^{14} C_i \left[\left[\frac{\exp(R_{mi})}{\left(\frac{1}{R_{mi}} \right) \log_{10} \left(\frac{1}{R_{mi}} \right)} \right] + \left[\sqrt{d_{mi}} \log_{10} \left(\frac{1}{d_{mi}} \right) \right] \right]}$$

Sir Occam will not
be happy !!

-
- Interpolation estimates can be corrected if knowledge exists about “no precipitation” conditions
 - This can be done before estimation or after estimation
 - Single Best Estimator, ANNs, Data mining principles, radar and satellite based (e.g. TRMM data) can be used for this purpose.
 - If the corrections are done prior to estimation, some computational effort can be saved and improvements in estimates can be achieved.

Integration of Spatial Interpolation and Association Rule Mining (data Mining)

-
- Data mining is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories.
 - *Association rule mining (ARM)* is mainly aimed at extracting interesting correlations, frequent patterns, associations or causal structures among data available in databases
 - ARM is carried out using an Apriori algorithm developed by Agarwal et al. (1993). Association rules take the form “if antecedent then consequent”.



WEKA: Waikato Environment for Knowledge Analysis

-
- Using the ARM based rules corrections are applied to the precipitation estimates provided by the spatial interpolation methods.
 - The ARM based rules can be translated to mathematical form

$$\text{if } (\bigcap (\theta_i = 0)), \text{ then } \theta_m^o = 0, \text{ else } \theta_m^o = \theta_m \quad \forall i, i \neq m$$

$$i \leq n-1$$

$$\alpha \geq \alpha_m$$

$$\beta \geq \beta_m$$

Rule	Description	α	β
1	Louisville = no rain \Rightarrow Lexington = no rain	0.55	0.89
2	Louisville = no rain, London = no rain \Rightarrow Lexington = no rain	0.50	0.95
3	Louisville = no rain, Berea = no rain \Rightarrow Lexington = no rain	0.50	0.95
4	Louisville = no rain, Somerset = no rain \Rightarrow Lexington = no rain	0.50	0.94

Mathematical Programming Models

Formulation

- Non-linear Mixed Integer programming formulation
 - Equality and Inequality (Hard and Soft Constraints)
 - Integers (binary variables)
- Solution algorithms
 - CONOPT and DICOPT (with mixed integers) within GAMS Environment
 - Genetic Algorithms
 - Simulated Annealing
- Combinatorial problem (due to existence of binary variables)
 - Computational intractability

Minimize $f(x, y)$

Subject to $h(x, y) = 0$

$g(x, y) \leq 0$

$x \in X \subseteq \Re$

$y \in Y$ integer

Minimize
$$\sum_{i=1}^{no} (\hat{\phi}_i^m - \phi_i^m)^2$$

Subject to:

$$\theta_i^m = \frac{\sum_{j=1}^{ns-1} w_{mj}^k \phi_i^j}{\sum_{j=1}^{ns-1} w_{mj}^k} \quad \forall i, no$$

$$0 \leq w_{mj}^k \leq 1 \quad \forall m, j$$

$$\frac{\sum_{i=1}^{no} |\hat{\phi}_i^m - \phi_i^m|}{no}$$

$$\theta_i^m = \frac{\sum_{j=1}^{ns-1} \lambda_j \theta_i^j w_{mj}^k}{\sum_{j=1}^{ns-1} \lambda_j w_{mj}^k} \quad \forall i, j, no$$


Subject to:

$$\sum_{j=1}^{ns-1} \lambda_j \leq np$$

np = upper limit on the number
of stations

λ = binary variable

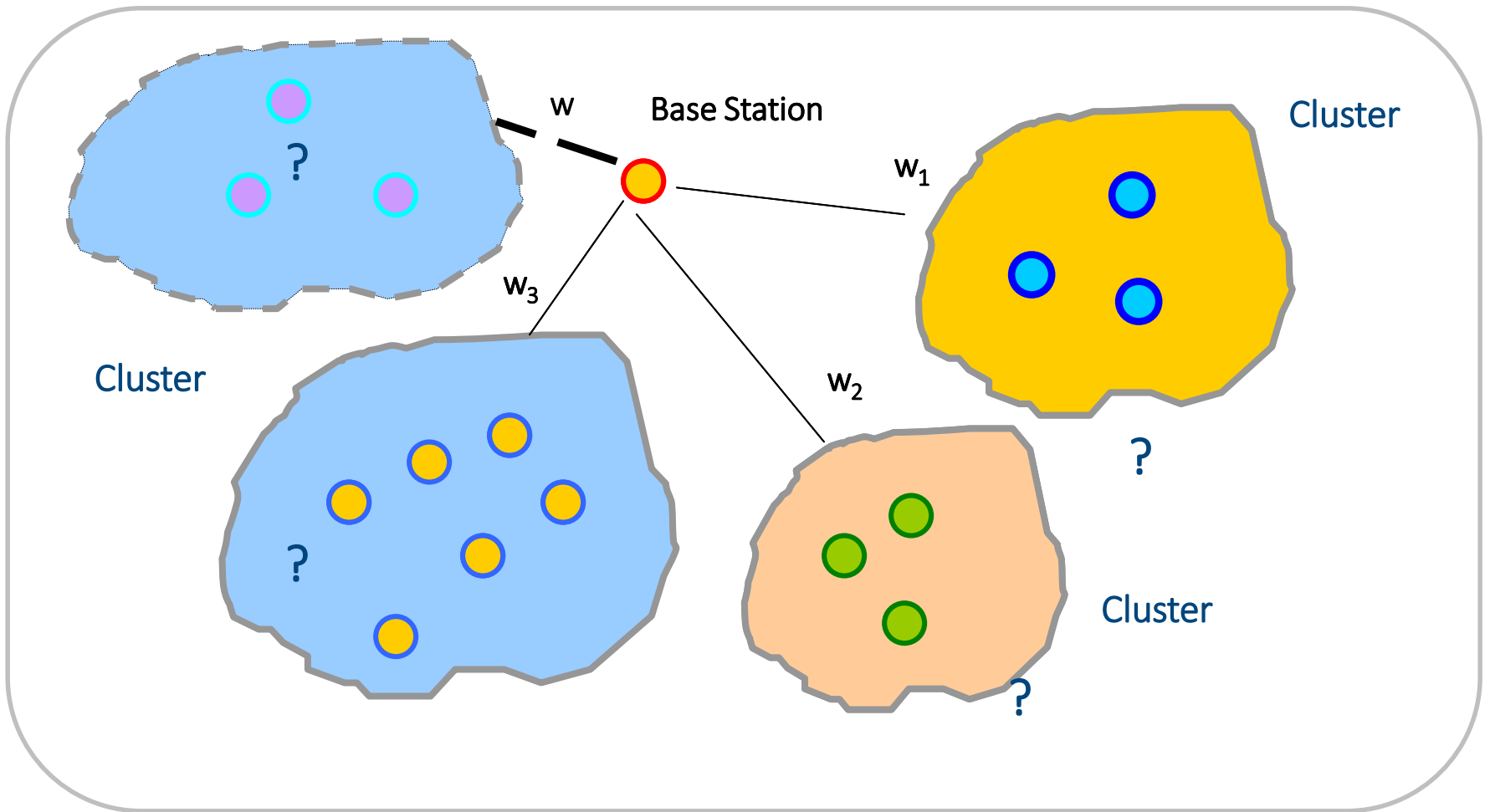
$$\theta_i^m = \frac{\sum_{j=1}^{ns-1} \left(\frac{\rho}{d_{mj}^n}\right)^2 \theta_i^j}{\sum_{j=1}^{ns-1} \left(\frac{\rho}{d_{mj}^n}\right)^2} \quad \forall i, no$$



Clusters

- Range
 - Development of local models (instead of one global model for the entire range of precipitation values)
 - Development of a Global model for encompassing all ranges of precipitation values
- Space
 - Development of cluster of gaging stations along with weights
 - Selection of these cluster in space and number and also in time.

-
- This formulation is referred to as “cluster approach” in which binary variables are used in a MINLP formulation to:
 - 1) select the cluster of stations which can be used for estimation process;
 - 2) define the number of clusters;
 - 3) select a discrete set of stations and
 - 4) define the maximum number of stations that can be used.



Model III

Minimize
$$\sum_{i=1}^{no} (\hat{\phi}_i^m - \phi_i^m)^2$$

Subject to:

$$\phi_i^m = \frac{\sum_{l=1}^{nc} [w_l (\sum_{j=1}^{N_l} \theta_j \lambda_{lj})]}{\sum_{l=1}^{nc} w_l}$$

$$\sum_{l=1}^{nc} \lambda_{lj} = 1 \quad \forall j$$

$$\sum_{j=1}^{N-1} \lambda_{lj} = N_l \quad \forall l$$

$$\sum_{l=1}^{nc} N_l = N - 1$$

Or

$$\sum_{l=1}^{nc} N_l \leq N - 1$$

l : cluster index

J :station index

N_l :number of
stations in
cluster " l "

$N-1$: Number
of stations

$$\sum_{l=1}^{nc} N_l = n_1^c + n_2^c + \dots n_{nc}^c$$

λ_{ij} : binary variable;

N-1: Number of stations used in interpolation;

nc: cluster size;

np: number of stations to be used;

w: weight.

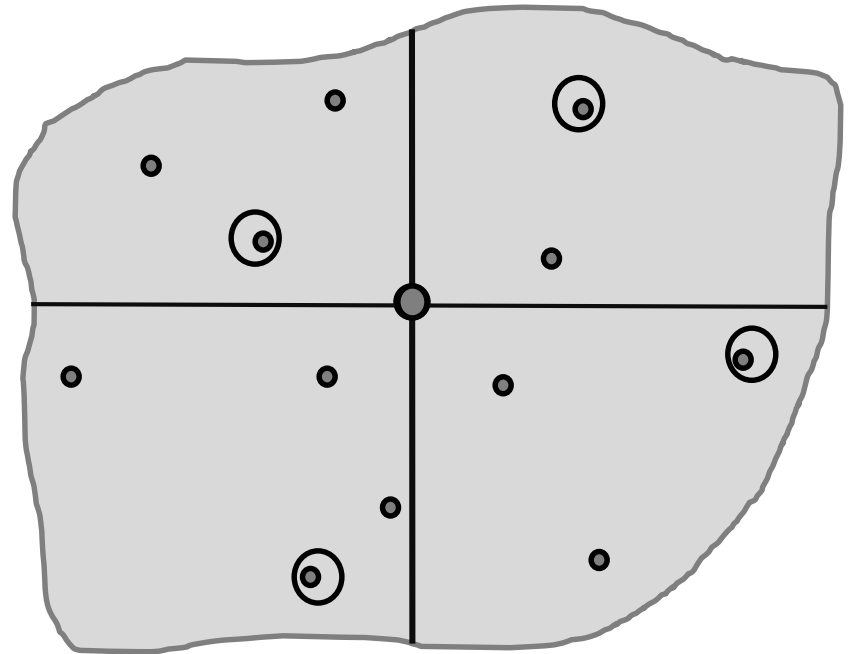
$n_1^c, n_2^c, \dots, n_{nc}^c$ are referred to as total number of stations in all the clusters

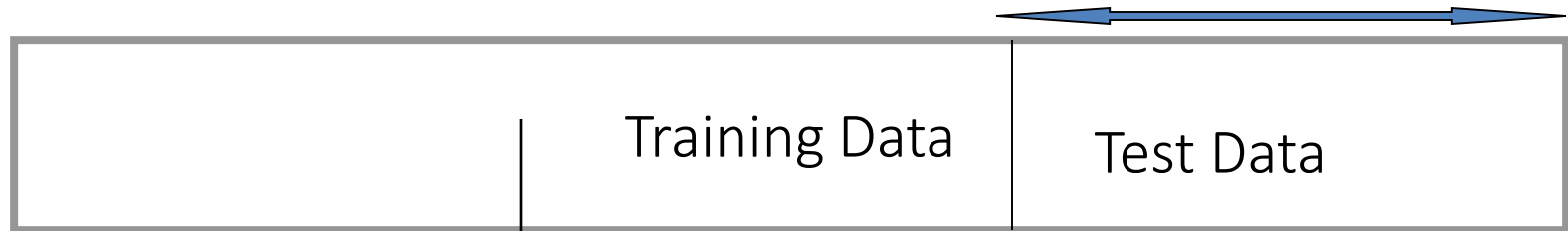
$$\varphi_i^m = \frac{\sum_{j=1}^{ns-1} \theta_i^j d_{m,j}^{-k} \lambda_j}{\sum_{j=1}^{ns-1} d_{m,j}^{-k} \lambda_j} \quad \forall i$$

$$\phi_i^m = \frac{\sum_{J=1}^4 \sum_{j \in J} \theta_i^j d_{m,j}^{-k} \lambda_j}{\sum_{J=1}^4 \sum_{j \in J} d_{m,j}^{-k} \lambda_j} \quad \forall i$$

$$\sum_{j \in J} \lambda_{j,J} \leq np_J \quad \forall J$$

$$\sum_{j \in J} \lambda_{j,J} = 1 \quad \forall J$$





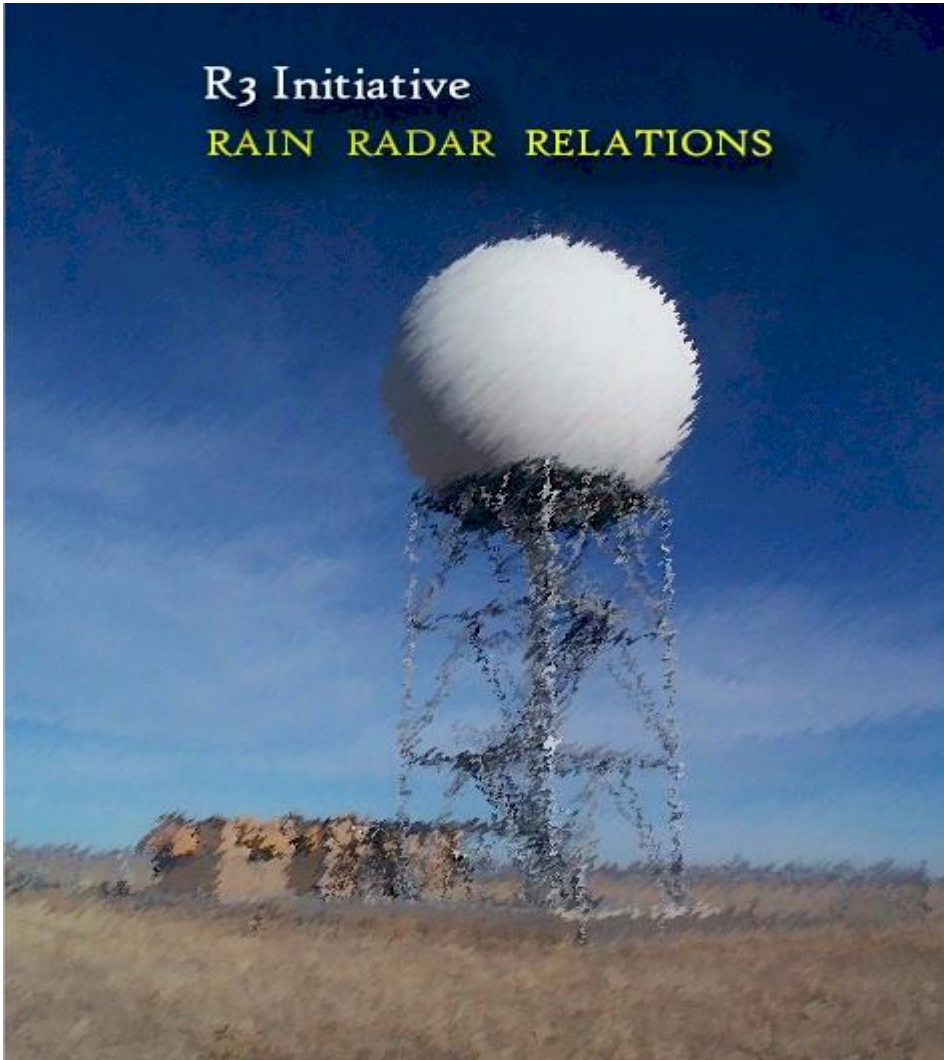
Weights
Variograms (type)
Connection weights in ANNs
Number of Stations
Number of Clusters
Cluster Size
Association rules

All the interpolation approaches require that the rain gage network remains un-changed for application in the test data period.

“Data, if tortured enough will
confess to any thing! “

-Anonymous

R₃ Initiative
RAIN RADAR RELATIONS



R₃ is a research initiative started at Florida Atlantic University (FAU)

Supported by the South Florida Water Management District

-
- Infilling of missing precipitation records is generally done using deterministic and stochastic interpolation methods with the help of rain gage data.
 - However, availability of radar based precipitation data with high spatial resolution can be an alternative to rain gage data. In many instances, radar data may be the only available data in a particular location

-
- Data derived from radar based precipitation estimates (i.e. NEXRAD data) can be used to estimate the missing precipitation values. However, the reliability of radar-based precipitation measurements is a contentious issue (Young et. al, 1999; Adler et al., 2001).
 - Radar rainfall estimates derived from conversion of reflectivity measurements are known to contain systematic errors, or bias, and other random errors or artifacts that limit the utility of radar rainfall.

-
- In-fill rainfall records based on radar-data data using a mathematical programming model to identify clusters of radar-data grids surrounding a rain gage.
 - Assess the utility of radar data for infilling purposes.
 - Investigation of spatial and temporal variability of clusters (identified by weights)

-
- Rainfall areas (or rain areas) are defined to represent the physical processes responsible for, or affecting, the genesis and morphology of rainfall processes near the coast and inland.
 - The delineation of these areas in south Florida is discussed in a study by Vieux (2006).
 - Rain areas may not coincide with watersheds

Simple Model Formulation

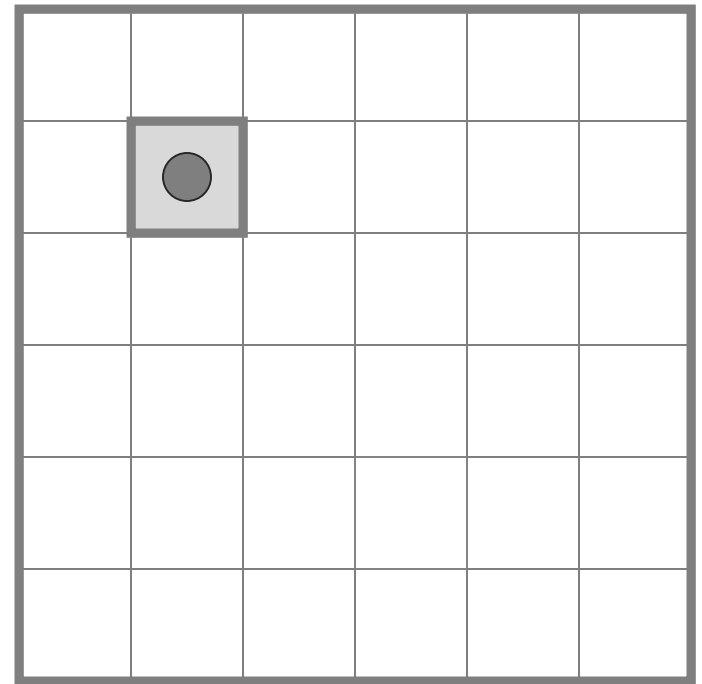
Model II

Infilling of missing precipitation data at a gage can be done using the NEXRAD based data available for that pixel (2km x 2km grid)

$$\theta_i^m = \phi_j$$

θ_i^m : estimated missing precipitation data value

ϕ_j Radar based precipitation data value



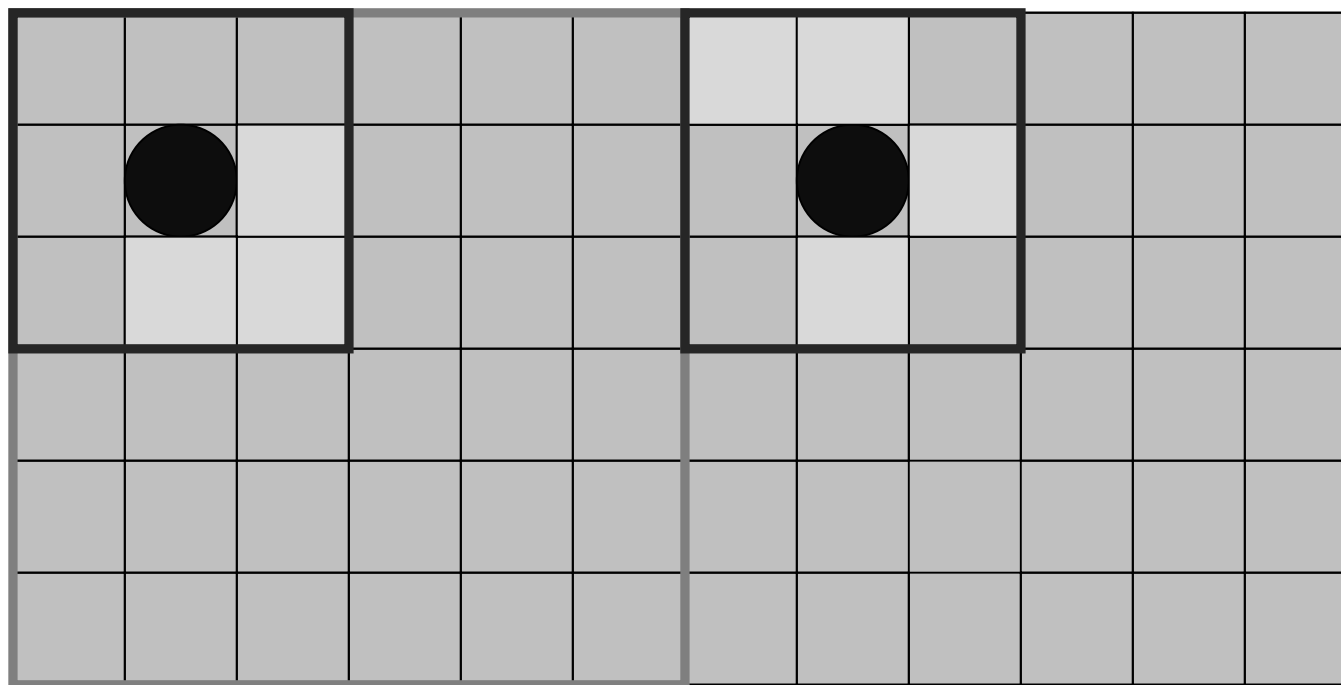
Clusters

Time



Dry

Wet



Space

Clusters : Selection and Number

Model I Formulation

Minimize
$$\sum_{i=1}^{no} (\hat{\theta}_i^m - \theta_i^m)^2$$

Subject to:

$$\theta_i^m = \sum_{j=1}^{nc} \lambda_j \phi_j w_j^k$$

$$\sum_{j=1}^{nc} \lambda_j = np$$

no : number of
days

i: day index

np : number of
cells

W: weight

$$\sum_{l=1}^{nc} \lambda_{lj} = 1 \quad \forall j$$

$$\sum_{j=1}^{N-1} \lambda_{lj} = N_l \quad \forall l$$

$$\sum_{l=1}^{nc} N_l = N$$

Or

$$\sum_{l=1}^{nc} N_l \leq N$$

l : cluster index

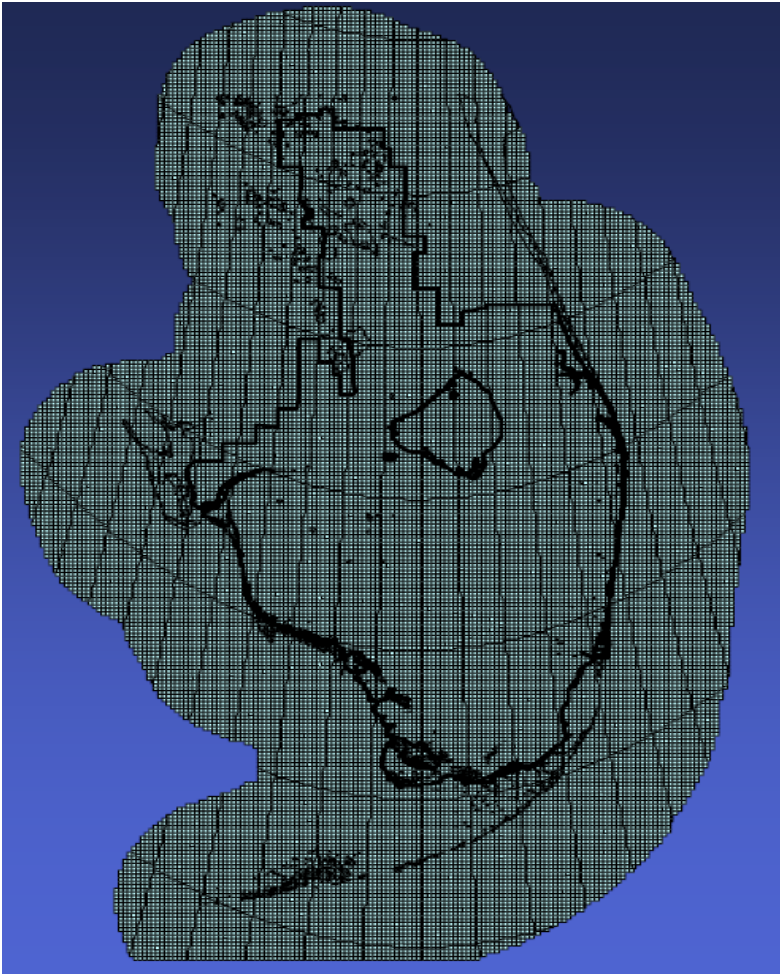
j : cell (pixel)
index

N_l : number of
pixels in cluster
" l "

N : Number of
pixels

Case Study Application

- The mathematical programming model for in-filling rainfall data is tested on a real-life case study system in an area where the gage adjusted RADAR-based (NEXRAD) rainfall data is available.
- The study areas selected are from Upper and Lower Kissimmee basins of south Florida. These areas formed the test-bed for the proposed approaches.
- NEXRAD rainfall data available over a 2 km by 2 km pixel size (grid size) in the study area along with rainfall data from 5 rain gages are used. Data from years 2002 -2004 are used for obtaining weights and testing of the models

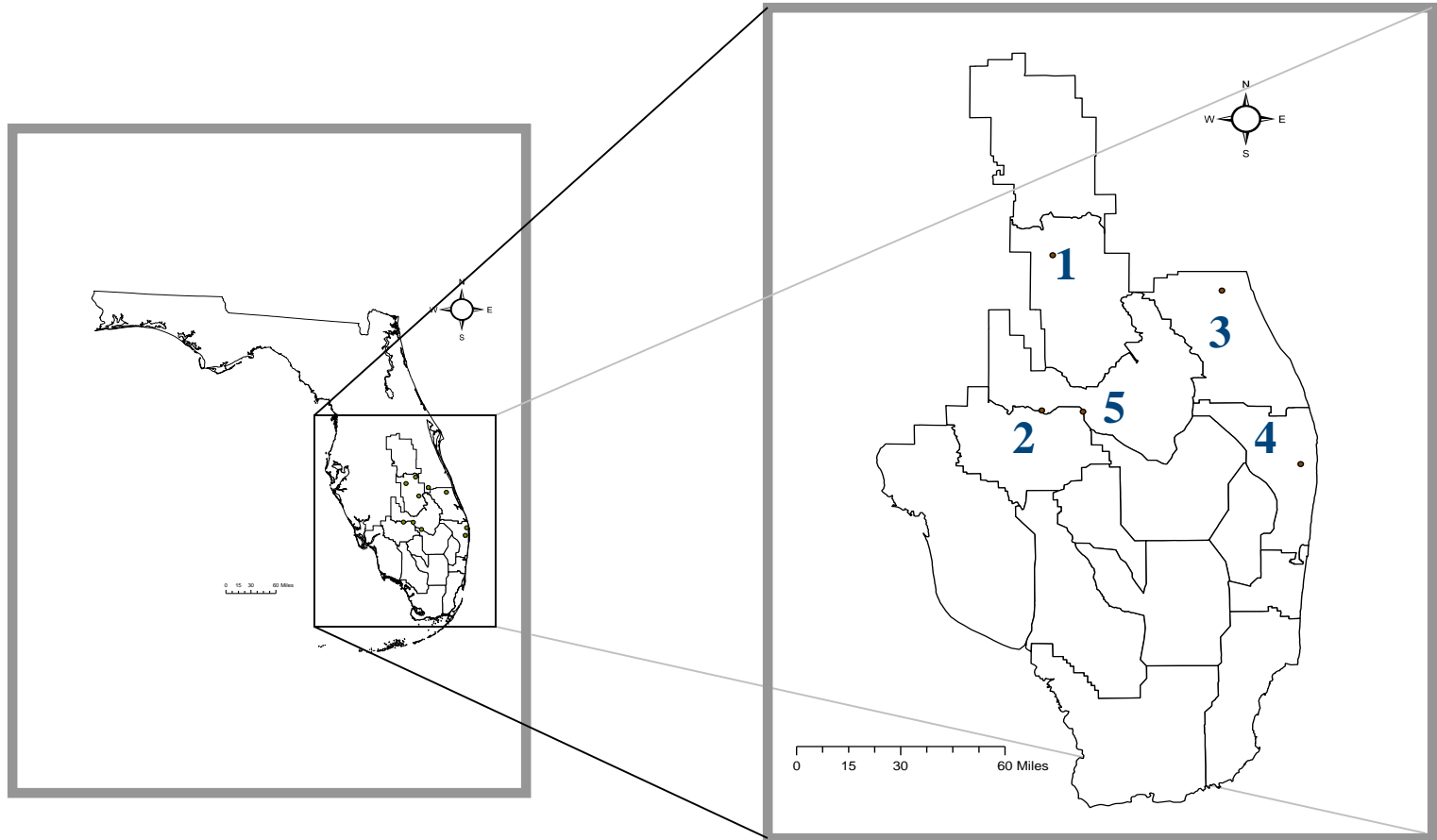


- 2 km X 2 km grid
- Base map (in state plane coordinates)
- 33,774 pixels (polygons)
- ~ 12,000 pixels within District
- Unique pixel id

Case Study Application : Clusters

- Five stations (referred to as cluster #1, 2, 3, 4 and 5) are used for the current study and these are:
 - Cluster # 1
 - (station: Avon Park, location: latitude:27 35 28; longitude: 81 31 07);
 - Cluster # 2
 - (station: Palmdale, location: latitude: 26 55 28; longitude: 81 18 50);
 - Cluster # 3
 - (station: S99, location: latitude:27 28 14; longitude: 80 28 18),
 - Cluster # 4
 - (station: WPB Airport, latitude:26 40 41; longitude: 80 06 35)
 - Cluster # 5
 - (station: CV5 latitude:26 55 10; longitude: 81 07 18).

Rain Areas and Selected Clusters

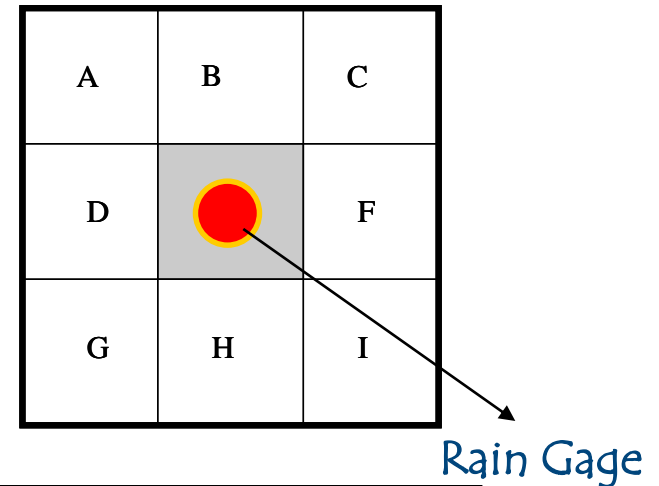


Results: Weights for cells

A 3 x 3 matrix pixels surrounding the gage are selected.

Weights are obtained based on

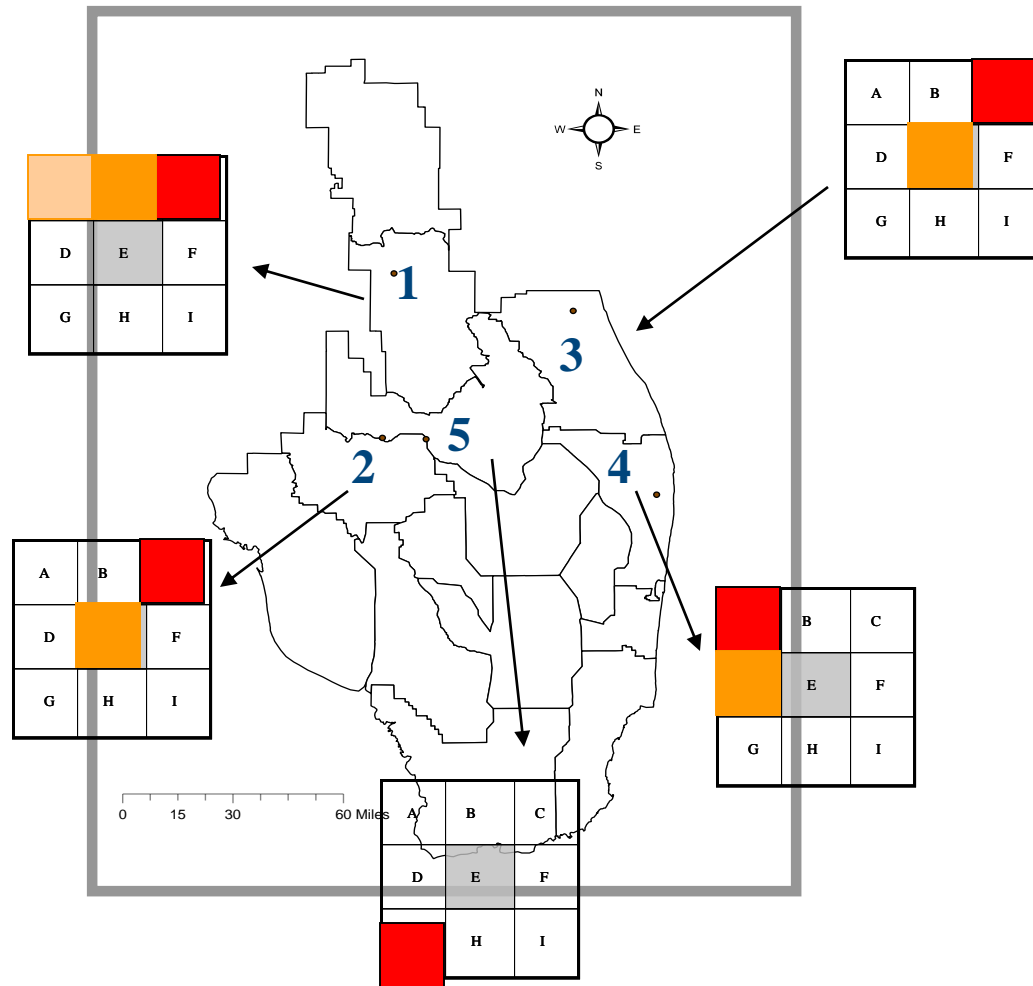
Model II formulation



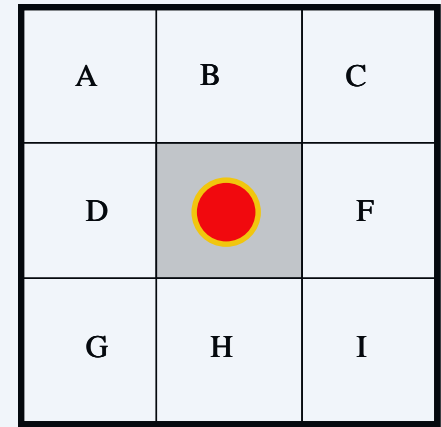
Cells Cluster	A	B	C	D	E	F	G	H	I
1	0.020	0.253	0.727	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.020	0.489	0.020	0.391	0.020	0.020	0.020	0.020
3	0.000	0.081	0.572	0.000	0.347	0.000	0.000	0.000	0.000
4	0.878	0.000	0.000	0.020	0.000	0.000	0.000	0.102	0.000
5	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000

Variation of
Weights in
The NEXRAD
pixels
surrounding the
gage.

All 9 pixel values
are allowed in the
Model I
formulation




-
- In almost all cases the pixel in which the rain gage was located did not receive the highest weight
 - On an average some cells had the lowest weight (example cell I)
 - Cell E had zero weight in case of clusters 4 and 5.



-
- An experiment was conducted where the number of cells that are allowed to participate in the infilling scheme is restricted

For Cluster # 1

Cells Restriction	A	B	C	D	E	F	G	H	I
1	0	0	1	0	0	0	0	0	0
2	0	1	1	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0
4	1	1	1	0	0	0	0	1	0
6	1	1	1	1	0	0	1	1	0

A	B	C
D		F
G	H	I

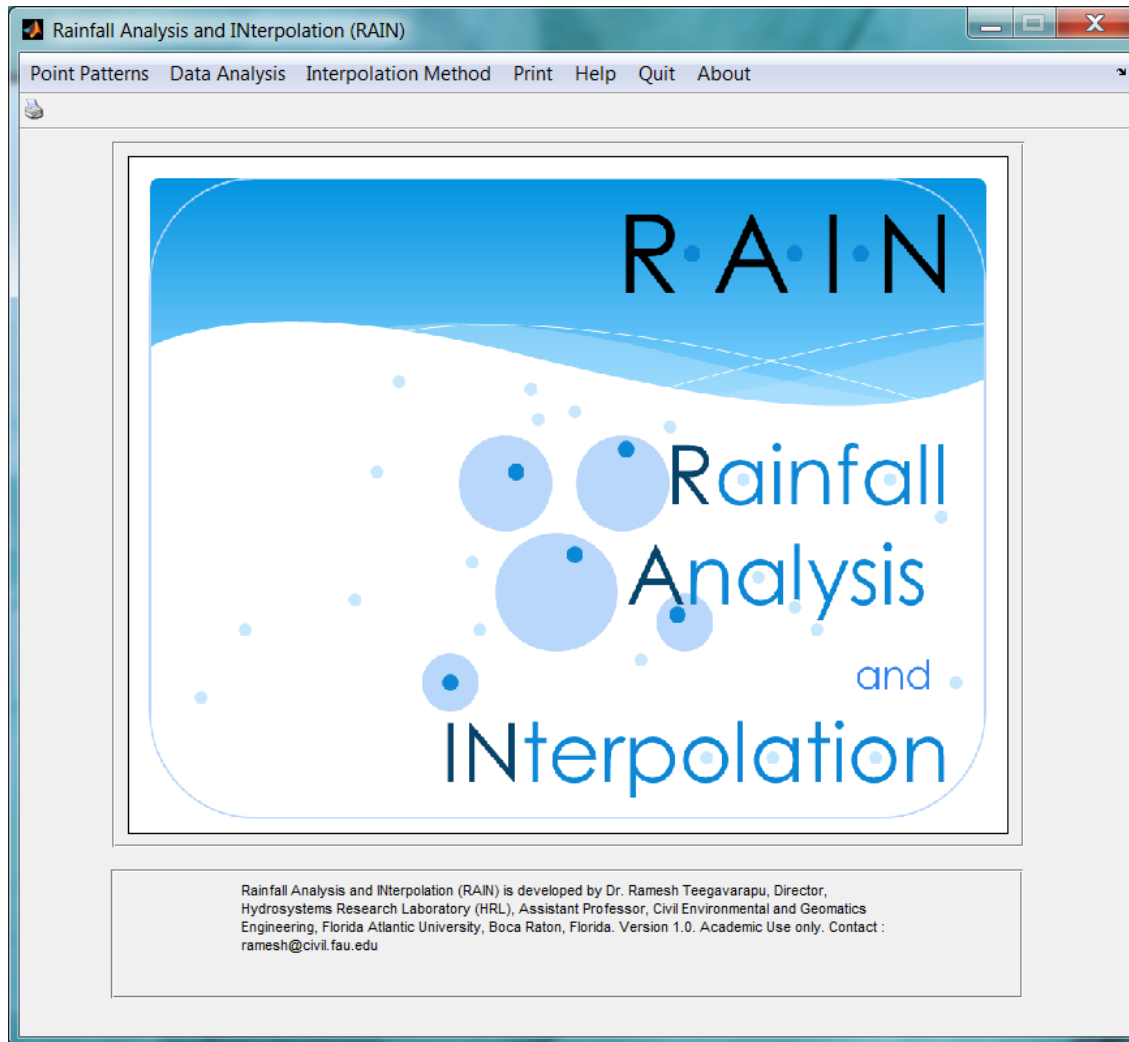
Software

(Developed at HRL, FAU)

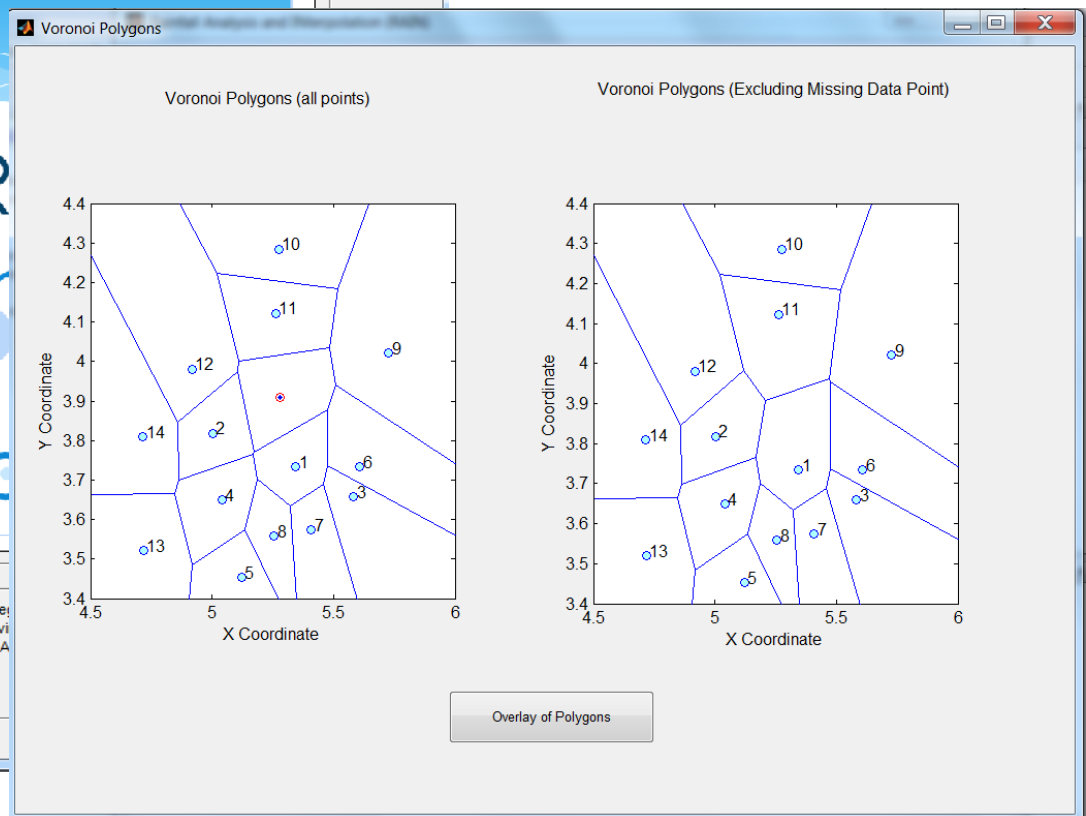
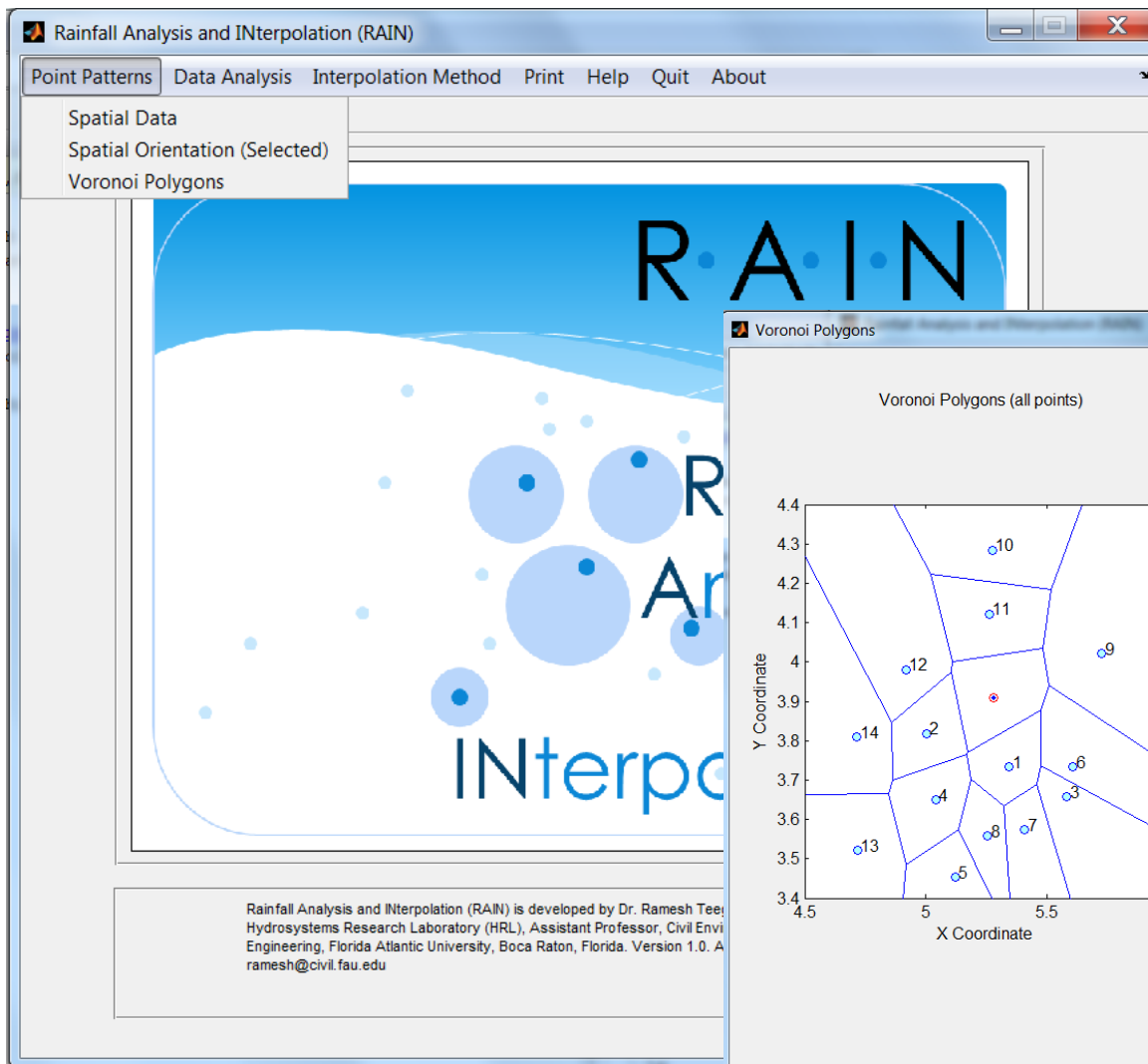
RAIN

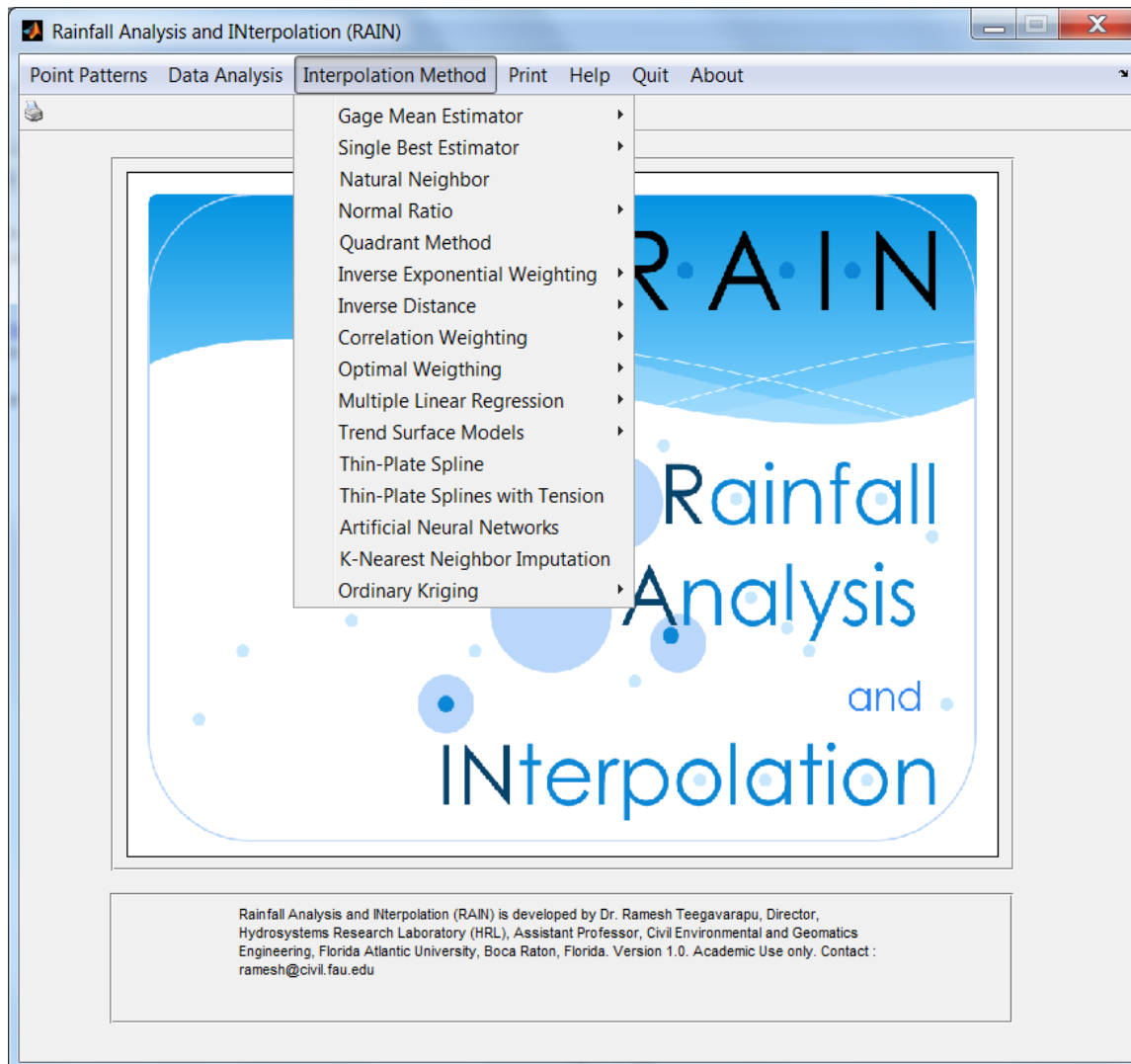
SEEP

R³



- A suite of deterministic and stochastic interpolation methods
- Inputs: Precipitation data at a location along with spatial coordinates of the observation stations in space.
- Outputs : Correlation matrices, precipitation data , estimated missing precipitation data, scatter graphs, residual plots, semi-variogram plots (kriging), Thiessen polygons, Negative value estimates, optimal parameters
- Jackknife Cross validation feature available.





Several different variants (over 40) of deterministic and stochastic Interpolation Methods

- Deterministic methods Including new methods
- Ordinary Kriging
 - Gaussian
 - Circular
 - Exponential
 - Spherical

Correlation Weighting Method (Nearest Neighbors & Exponents)

Total Control Points

14

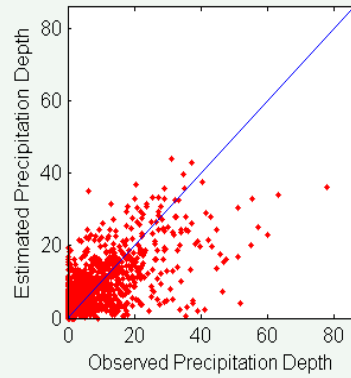
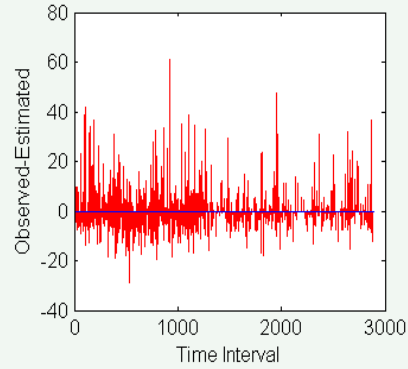
Neighbors

9

Exponent

4

Calculate

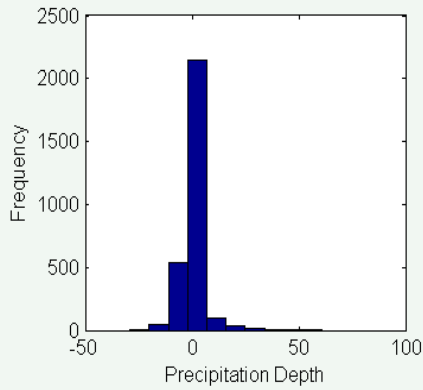


Correlation Coefficient

0.739621

Absolute Error

6917.1



Mean Absolute Error

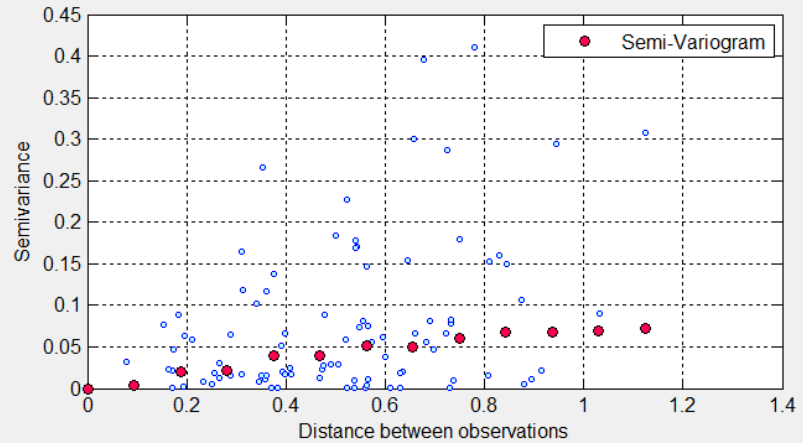
Root Mean Squared Error

Variogram Cloud and Semivariogram

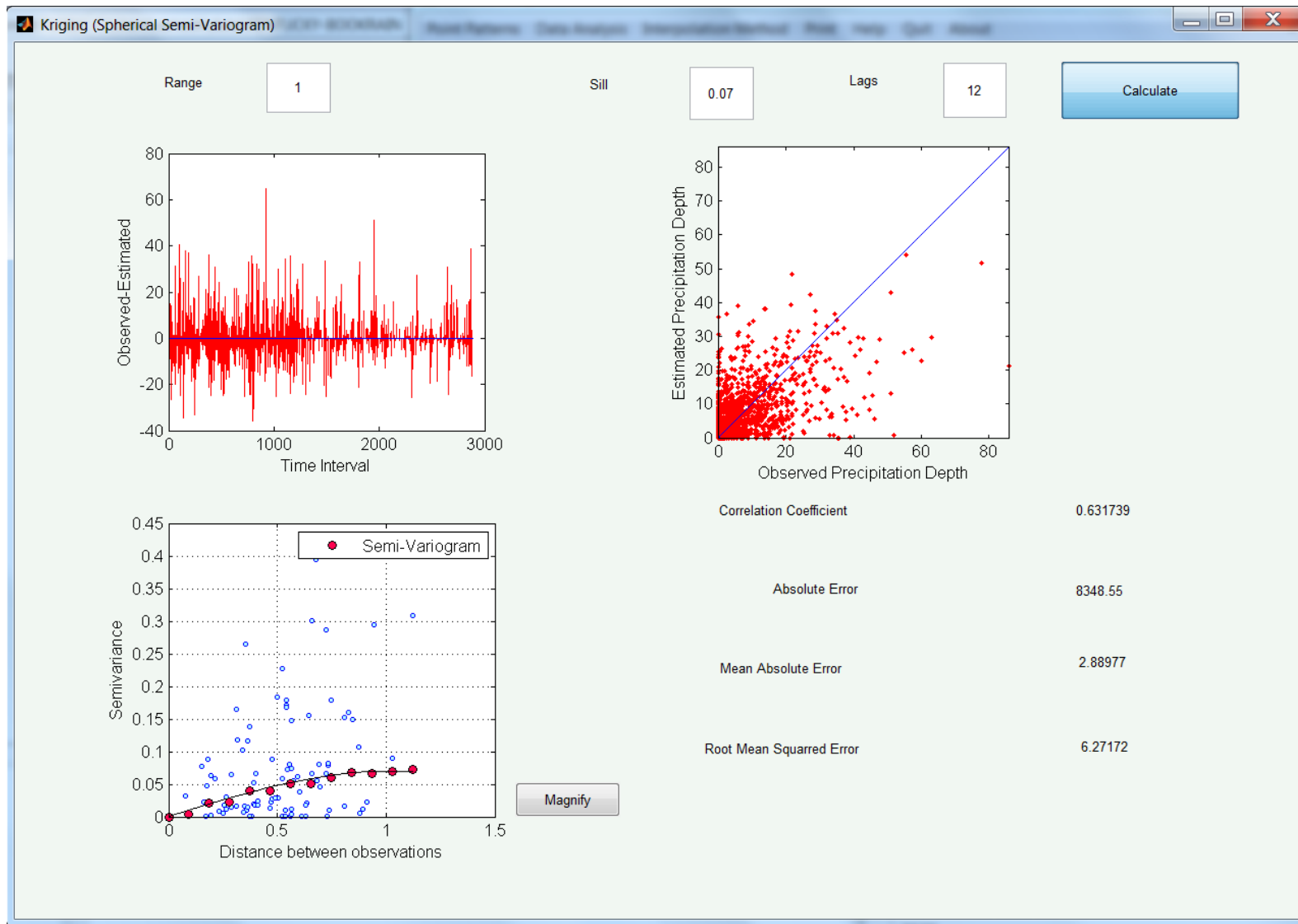
Number of Lags

12

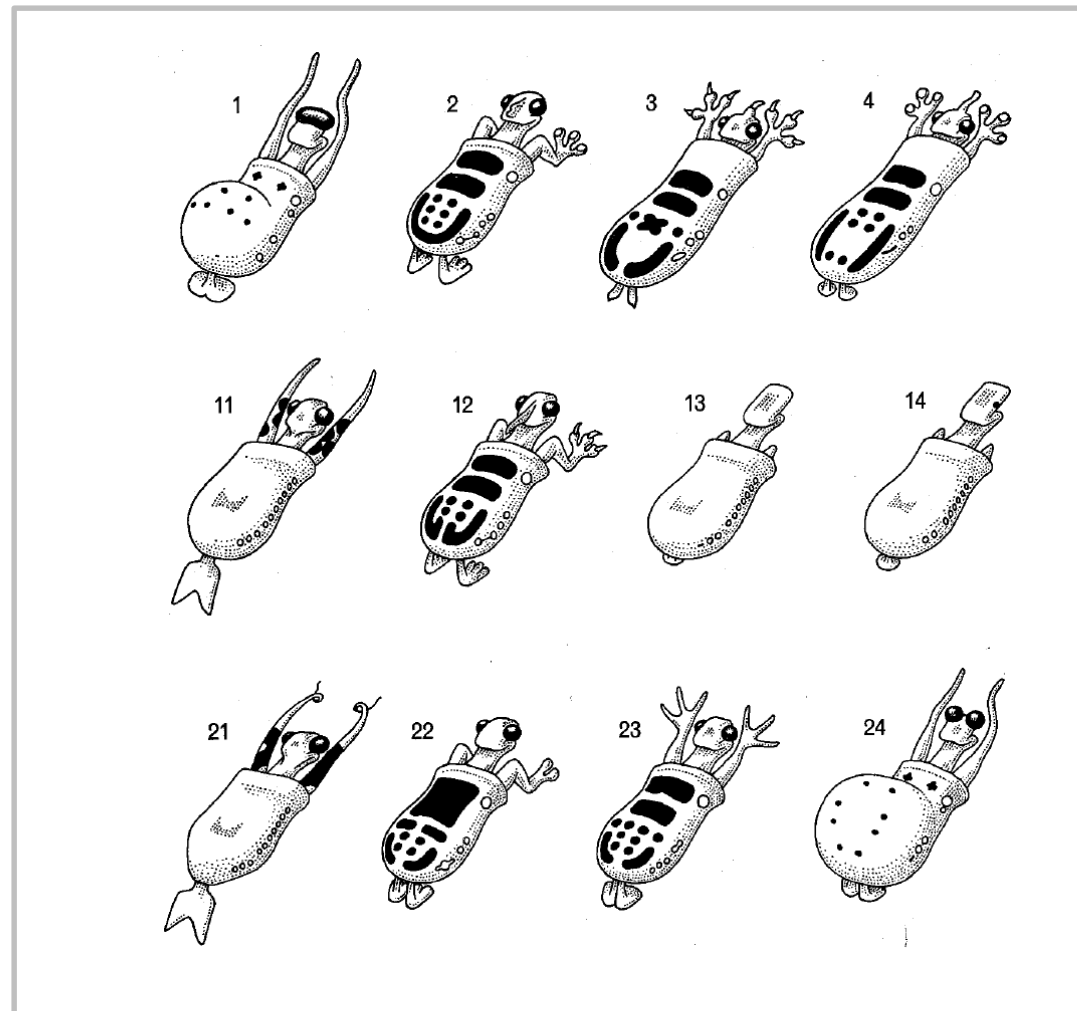
Calculate



Ordinary Kriging



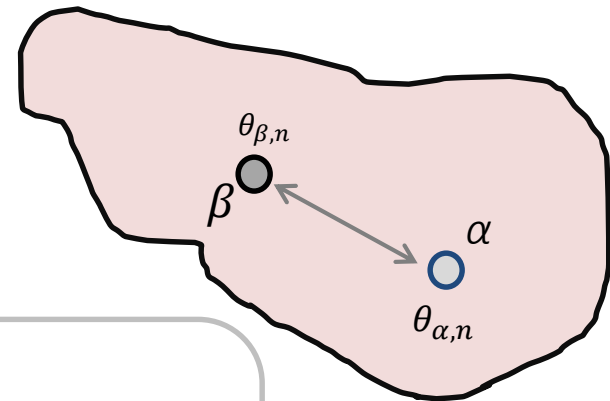
Proximity Metric-based Methods for Filling Missing Data



-
- **Euclidean distance** forms a central idea to many spatial interpolation methods.
 - Moving away from the concepts of using Euclidean distances as surrogates for spatial correlations, **numerical distances strictly based on observations** can be developed.
 - Ahrens (2006) identified one such distance referred to as statistical distance based on common variance of precipitation time series and reported improvements in estimation compared to geographical distance-based methods.
 - If rainfall surface or field estimation is **not of interest** and only estimation of missing data **at one single location in a region is essential**, methods which rely directly on observed data should be well suited for interpolation schemes.

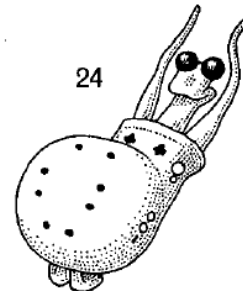
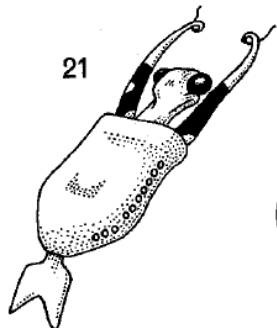
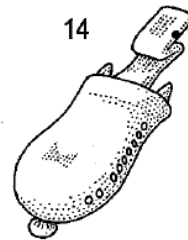
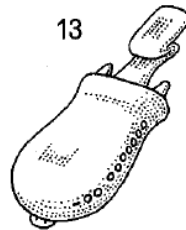
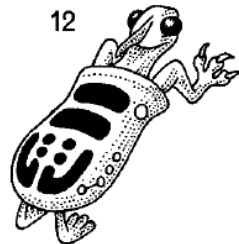
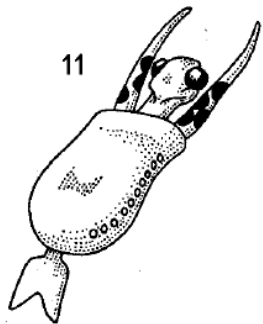
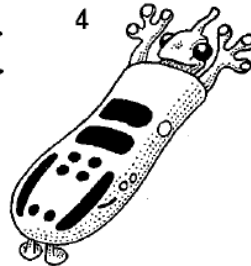
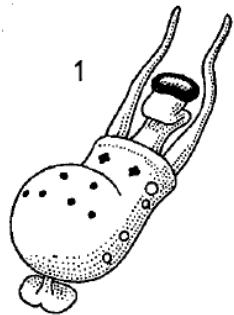
-
- Can we use proximity measures (not based on Euclidean distances) for developing spatial interpolation methods for estimation of missing precipitation data at a site ?
 - Are the methods better when used in local or global interpolation modes ?
 - Can we classify the sites into several clusters and use clusters as virtual stations for estimation of missing precipitation data ?
 - Can we classify the sites into several clusters and use one representative site within each cluster selected by a pre-specified criterion or optimization ?
 - Can we use division of spatio-temporal precipitation data into pre-specified classes to improve spatial interpolation via nearest neighbor classification and by developing local models ?

-
- Distance measures based on observations (θ_β and θ_α) at two rain gaging stations, β and α , can be defined as real-valued functions. The functions are referred to as *distance metrics* if they satisfy several conditions given by the following inequalities.



- $d_{\beta,\alpha} \geq 0$ (non-negativity)
- $d_{\beta,\alpha} = 0$ if and only if $\theta_{\beta,n} = \theta_{\alpha,n} \quad \forall n$ (Equality)
- $d_{\beta,\alpha} = d_{\alpha,\beta}$ (Commutativity)
- $d_{\beta,\omega} \leq d_{\beta,\alpha} + d_{\alpha,\omega}$ (Triangular Inequality)

-
- The grouping by **numerical methods** of taxonomic units based on their character states (Sneath and Sokal)
 - Taxonomy is the theoretical **study of classification**, including its bases, principles, procedures, and rules (Simpson),
 - **Numerical taxonomy** provides methods that are **objective, explicit, and repeatable**, and is based on the ideas first put forward by Adanson (1963).
 - Ideal taxonomy is composed of information-rich taxa ("**taxon**," plural "**taxa**," is the abbreviation for taxonomic group of any nature or rank (Lam) :
 - Incorporates **many features** as possible
 - Every **character is of equal weight**
 - Overall similarity between any two entities is a function of the **similarity of the many characters on which the comparison is based**.



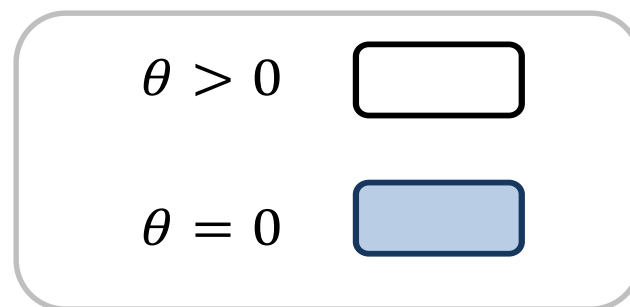
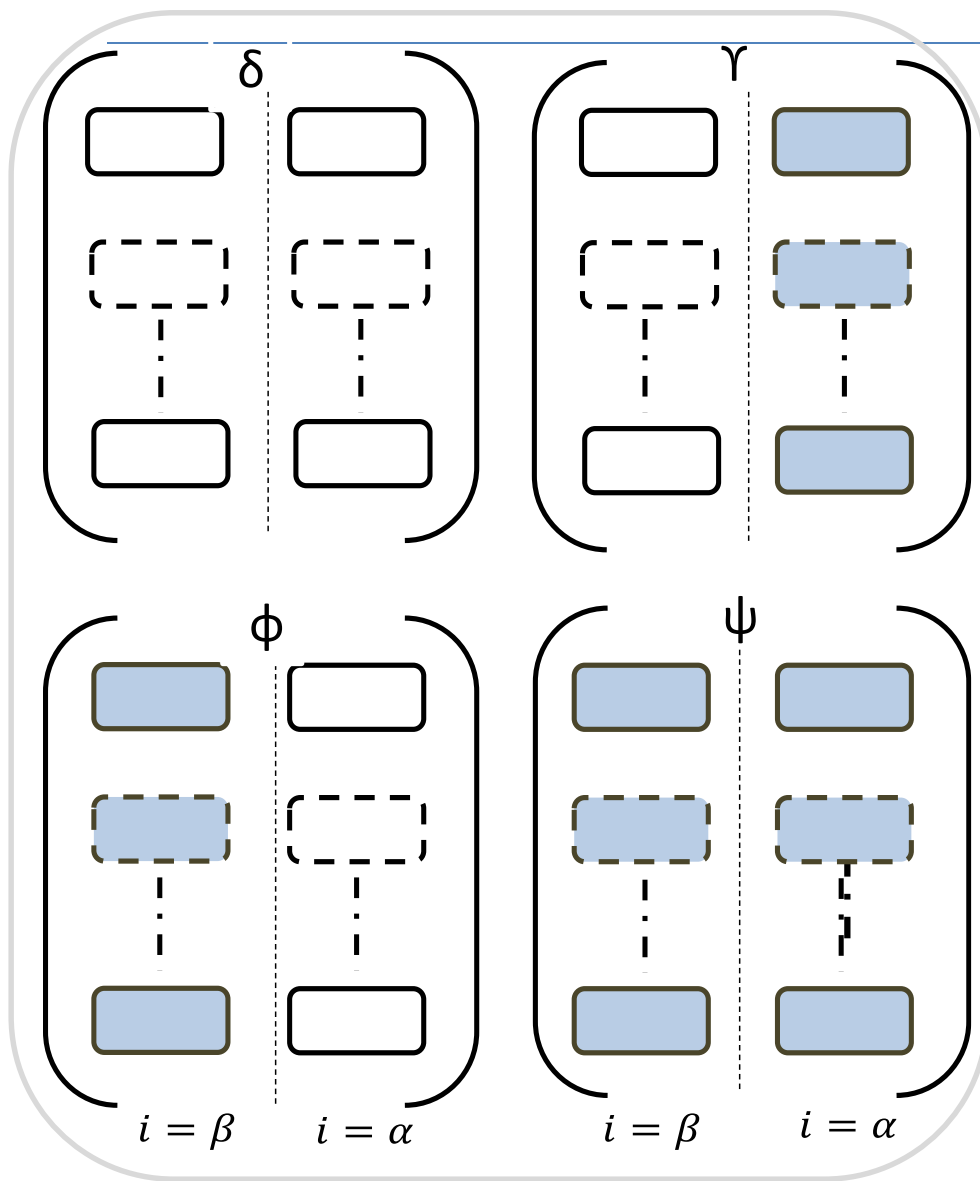
Imaginary Animals

“Caminalcules”

Created by J. H. Camin
Introduced by Sokal (1966)
in an original article in Nature.

- Classification: grouping of unordered items.
- Identification: Allocation of unidentified objects to different classes.

-
- Distance Measures can be **binary** or **real**
 - Measures need time consistent, gap free serial datasets at two sites or for two entities of interest
 - Binary distance measures are calculated for datasets that are in binary form.
 - Observations at two rain gage locations can be used for calculation of real measures or binary transformed observations can be used for calculation of binary measures

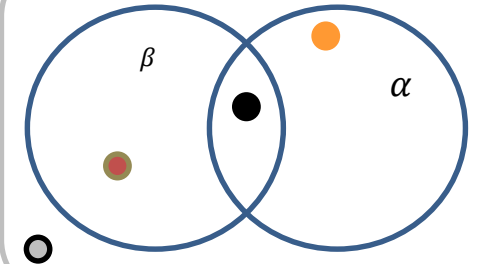


		θ_α	
		> 0	$= 0$
θ_β	> 0	δ	γ
	$= 0$	ϕ	ψ

if $(\theta_{j,n} > \theta_{th})$, then $\theta_{j,n}^* = 1$, else $\theta_{j,n}^* = 0 \quad \forall l, n$

Binary transformation
using a threshold (θ_{th}) , $\theta_{j,n}^*$ is
binary transformed value

if $(\theta_{\beta,n}^* = 1, \theta_{\alpha,n}^* = 1)$, then $\delta_n = 1, \quad \beta \in j \quad \forall \beta, n$ ●
if $(\theta_{\beta,n}^* = 1, \theta_{\alpha,n}^* = 0)$, then $\gamma_n = 1, \quad \beta \in j \quad \forall \beta, n$ ●●
if $(\theta_{\beta,n}^* = 0, \theta_{\alpha,n}^* = 1)$, then $\phi_n = 1, \quad \beta \in j \quad \forall \beta, n$ ●●
if $(\theta_{\beta,n}^* = 0, \theta_{\alpha,n}^* = 0)$, then $\psi_n = 1, \quad \beta \in j \quad \forall \beta, n$ ●●



$$\gamma_{\beta,\alpha} = \sum_{n=1}^{no} \gamma_n$$

$$\psi_{\beta,\alpha} = \sum_{n=1}^{no} \psi_n$$

$$\phi_{\beta,\alpha} = \sum_{n=1}^{no} \phi_n$$

$$\delta_{\beta,\alpha} = \sum_{n=1}^{no} \delta_n$$

Counts are obtained for
the conditions using the
transformed
rain gage observations
 n : index for time
intervals.

Proximity Metrics (Real)

$$d_{\beta,\alpha} = \sum_{n=1}^{no} \sqrt{(\theta_{\beta,n} - \theta_{\alpha,n})^2}$$

Euclidean

$$d_{\beta,\alpha} = \sqrt[\lambda]{\sum_{n=1}^{no} |\theta_{\beta,n} - \theta_{\alpha,n}|^\lambda}$$

Minkowski

$$d_{\beta,\alpha} = \sum_{n=1}^{no} (\theta_{\beta,n} - \theta_{\alpha,n})^2$$

Sq. Euclidean

$$d_{\beta,\alpha} = \sqrt{\frac{\sum_{n=1}^{no} \omega_n d_n^2}{\sum_{n=1}^{no} \omega_n}}$$

Gower

$$d_n = \frac{|\theta_{\beta,n} - \theta_{\alpha,n}|}{\tau_n} \quad \forall n$$

$$d_{\beta,\alpha} = \sum_{n=1}^{no} |\theta_{\beta,n} - \theta_{\alpha,n}|$$

Manhattan

$$d_{\beta,\alpha} = \max |\theta_{\beta,n} - \theta_{\alpha,n}|$$

$\forall n$

Maximum

$$d_{\beta,\alpha} = 1 - \frac{\sum_{n=1}^{no} \theta_{\beta,n} \theta_{\alpha,n}}{\sqrt{\sum_{n=1}^{no} \theta_{\beta,n}^2 \sum_{n=1}^{no} \theta_{\alpha,n}^2}}$$

Cosine

$$d_{\beta,\alpha} = \sum_{n=1}^{no} \frac{|\theta_{\beta,n} - \theta_{\alpha,n}|}{(|\theta_{\beta,n}| + |\theta_{\alpha,n}|)}$$

Canberra

$$d_{\beta,\alpha} = 1 - \rho_{\beta,\alpha}$$

Correlation

$$d_{\beta,\alpha} = \sqrt{(\theta_{\beta,n} - \theta_{\alpha,n})S^{-1}(\theta_{\beta,n} - \theta_{\alpha,n})^T}$$

Mahalanobis

$$d_{\beta,\alpha} = 1 - \frac{\delta_{\beta,\alpha} + \gamma_{\beta,\alpha}}{\delta_{\beta,\alpha} + \gamma_{\beta,\alpha} + \phi_{\beta,\alpha} + \psi_{\beta,\alpha}}$$

Simple Matching

$$d_{\beta,\alpha} = \frac{\gamma_{\beta,\alpha} + \phi_{\beta,\alpha}}{\delta_{\beta,\alpha} + \gamma_{\beta,\alpha} + \phi_{\beta,\alpha}}$$

Jaccard

$$d_{\beta,\alpha} = \frac{2(\gamma_{\beta,\alpha} + \phi_{\beta,\alpha})}{\delta_{\beta,\alpha} + 2(\gamma_{\beta,\alpha} + \phi_{\beta,\alpha}) + \psi_{\beta,\alpha}}$$

Rogers and Tanimoto

$$d_{\beta,\alpha} = 1 - \frac{\delta_{\beta,\alpha}}{\delta_{\beta,\alpha} + \gamma_{\beta,\alpha} + \phi_{\beta,\alpha} + \psi_{\beta,\alpha}}$$

Russell & Rao

$$d_{\beta,\alpha} = \frac{\gamma_{\beta,\alpha} \phi_{\beta,\alpha}}{\delta_{\beta,\alpha} \psi_{\beta,\alpha} + \gamma_{\beta,\alpha} \phi_{\beta,\alpha}}$$

Yule

$$d_{\beta,\alpha} = \frac{\gamma_{\beta,\alpha} + \phi_{\beta,\alpha}}{2\delta_{\beta,\alpha} + \gamma_{\beta,\alpha} + \phi_{\beta,\alpha}}$$

Dice

$$d_{\beta,\alpha} = \frac{1}{2} - \frac{\delta_{\beta,\alpha} \psi_{\beta,\alpha} + \phi_{\beta,\alpha} \gamma_{\beta,\alpha}}{2\sqrt{(\delta_{\beta,\alpha} + \gamma_{\beta,\alpha})(\delta_{\beta,\alpha} + \phi_{\beta,\alpha})(\gamma_{\beta,\alpha} + \psi_{\beta,\alpha})(\phi_{\beta,\alpha} + \psi_{\beta,\alpha})}}$$

Pearson

$$d_{\beta,\alpha} = \frac{\delta_{\beta,\alpha} + \psi_{\beta,\alpha}}{\delta_{\beta,\alpha} + \gamma_{\beta,\alpha} + \phi_{\beta,\alpha} + \psi_{\beta,\alpha}}$$

Sokal-Michener

$$d_{\beta,\alpha} = \frac{2\gamma_{\beta,\alpha} + 2\phi_{\beta,\alpha} + \psi_{\beta,\alpha}}{\delta_{\beta,\alpha} + 2\gamma_{\beta,\alpha} + 2\phi_{\beta,\alpha} + \psi_{\beta,\alpha}}$$

Kulzinksy

$$d_{\beta,\alpha} = \gamma_{\beta,\alpha} + \phi_{\beta,\alpha}$$

Hamming

Minimize $\sum_{n=1}^{no} |\theta'_{\alpha,n} - \theta_{\alpha,n}|$

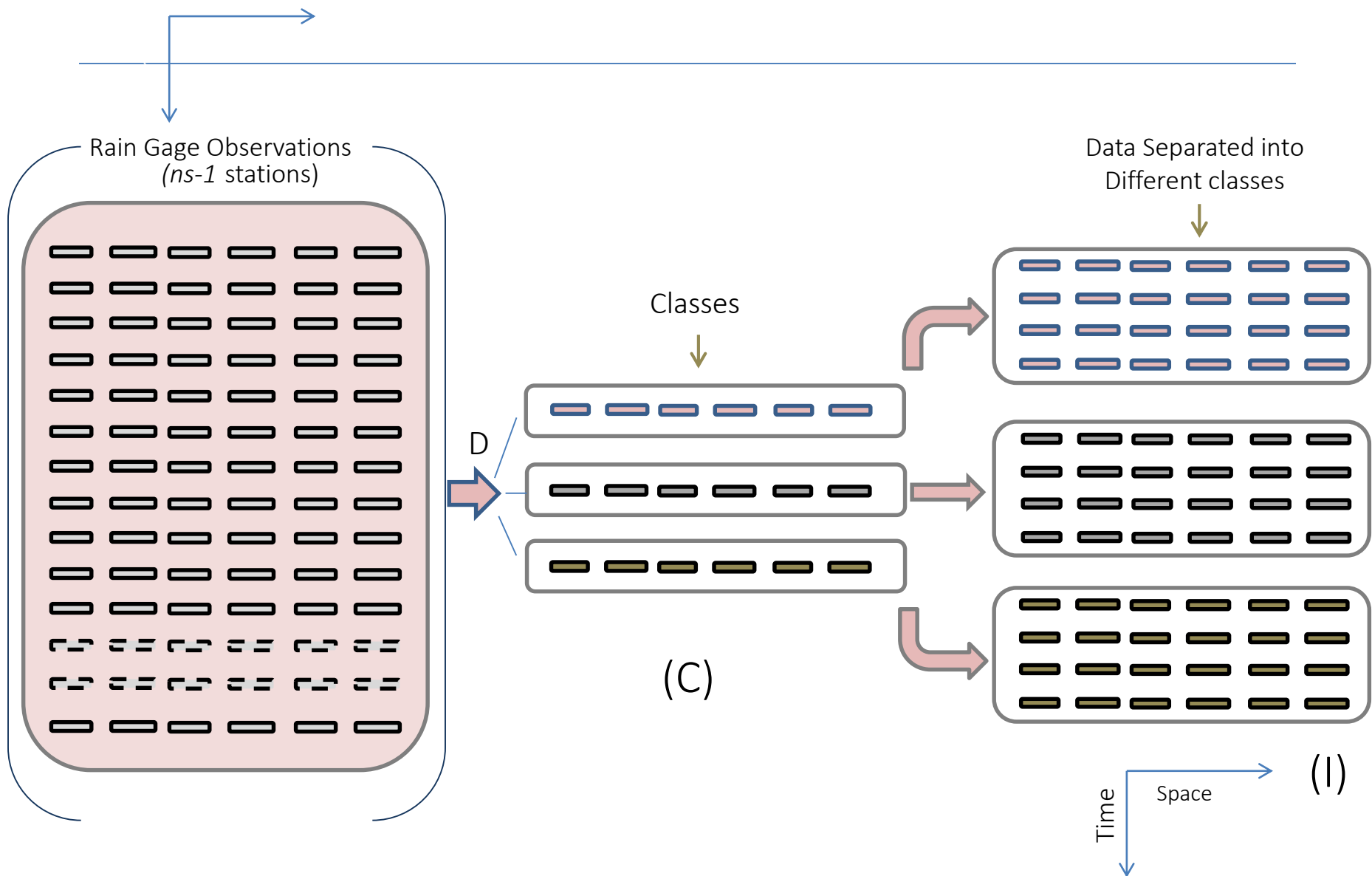
Subject to:

$$\theta'_{\alpha,n} = \frac{\sum_{j=1}^{ns-1} w_{\alpha,j}^k \theta_{j,n}}{\sum_{j=1}^{ns-1} w_{\alpha,j}^k} \quad \forall n \quad j \notin \alpha$$

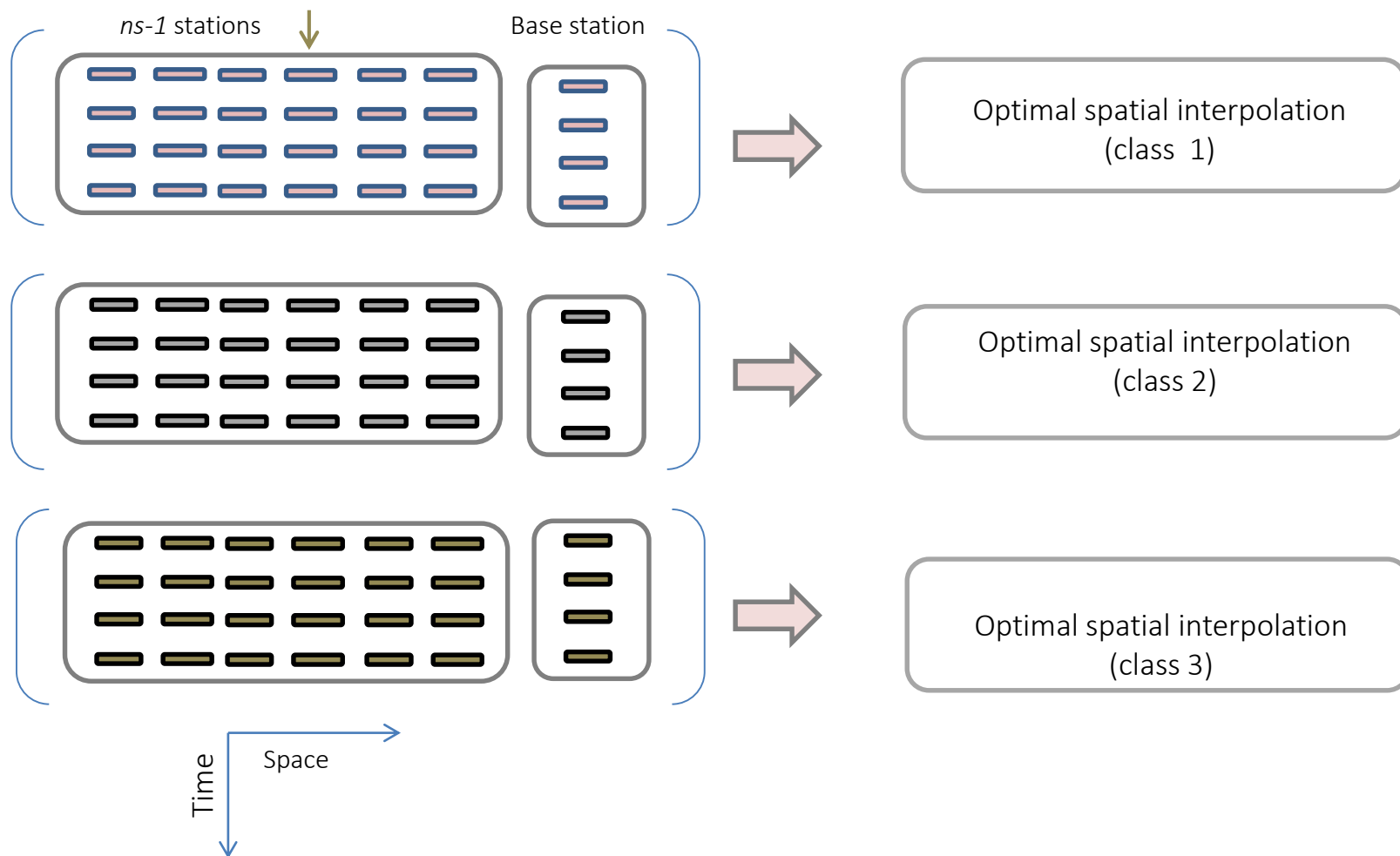
$$w_{\alpha,j} = \frac{1}{d_{j,\alpha}} \quad j \notin \alpha, \quad j \in ns - 1$$

Formulation
is used to optimize
exponent
associated with
the inverse of the
distances
obtained from
proximity
metrics

-
- The objective was to see if classification of precipitation data into several groups to develop separate (i.e., local) optimization models would yield better results compared to those from a global model.
 - Arbitrary definitions of classes have yielded inferior results in the previous study (Teegavarapu, 2012).
 - Initially, the k -nearest-neighbor method is used as a classifier to group spatial and temporal precipitation data into several pre-defined classes.



Data Separated into
Different classes



- Precipitation data

$$X = \begin{pmatrix} \theta_{1,1} & \cdots & \theta_{1,ns-1} \\ \vdots & \ddots & \vdots \\ \theta_{no,1} & \cdots & \theta_{no,ns-1} \end{pmatrix}$$

Spatial-temporal observations
no: number of observations
ns-1: number of stations excepting the base station

- Pre-defined classes

$$X_c = \begin{pmatrix} \theta_{1,1}^\circ & \cdots & \theta_{1,ns-1}^\circ \\ \vdots & \ddots & \vdots \\ \theta_{c,1}^\circ & \cdots & \theta_{c,ns-1}^\circ \end{pmatrix}$$

Classes
c: number of classes
ns-1: number of stations excepting the base station

- Distance Metric Calculation $D = f(X, X_c)$

Distance metric (*D*) to classify data into different classes

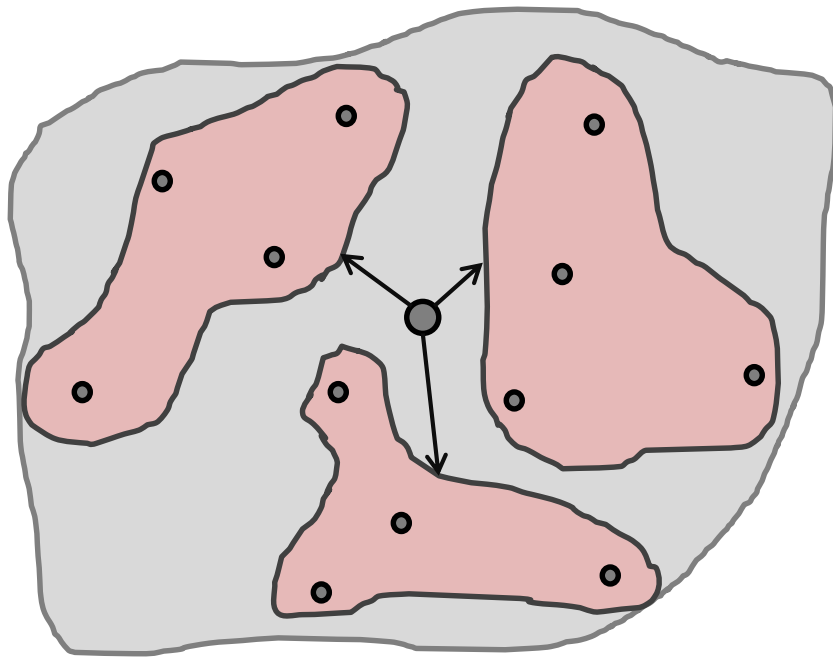
- Missing Data

$$\theta'_{\alpha,n} = \frac{\sum_{j=1}^{ns-1} \rho_{\alpha,j,c} \theta_{j,nc}}{\sum_{j=1}^{ns-1} \rho_{\alpha,j,c}} \quad \forall nc \in n, j \notin \alpha$$

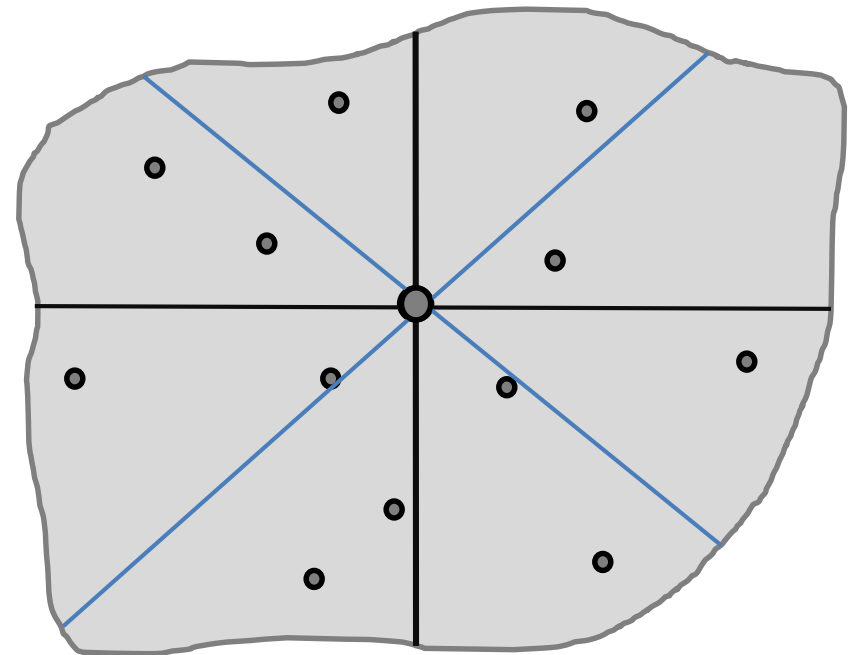
Missing data estimation using Correlation Weighting

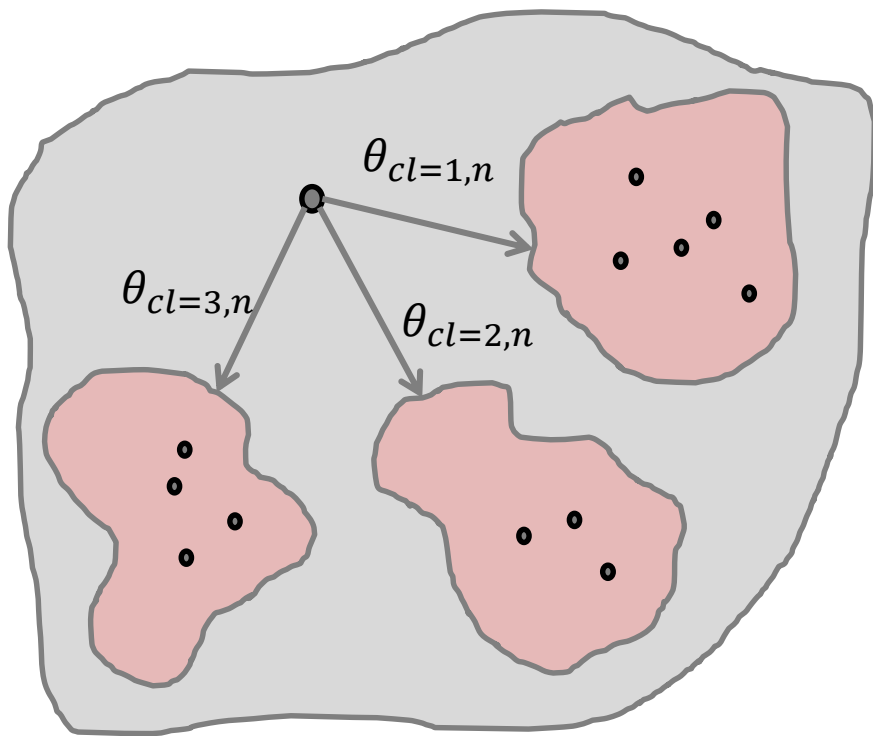
-
- Clustering of observations is quite common in spatial ecology (Dungan et al., 2002; Fortin and Dale, 2005), geosciences and data mining applications (Jain and Dubes, 1988).
 - Clustering can be used to create one “virtual” station out of a number of stations (i.e., cluster) where only one weight is attached to this station (i.e. group of stations).
 - Grouping (clustering) of stations can be achieved by k-means clustering algorithms where proximity (distance-based) metrics are used to define the clusters and the members belonging to each cluster.
 - Clustering can be achieved by optimization using binary variable-mixed integer nonlinear programming formulation (Teegavarapu, 2012).

Fixed Spatial Division for Clusters



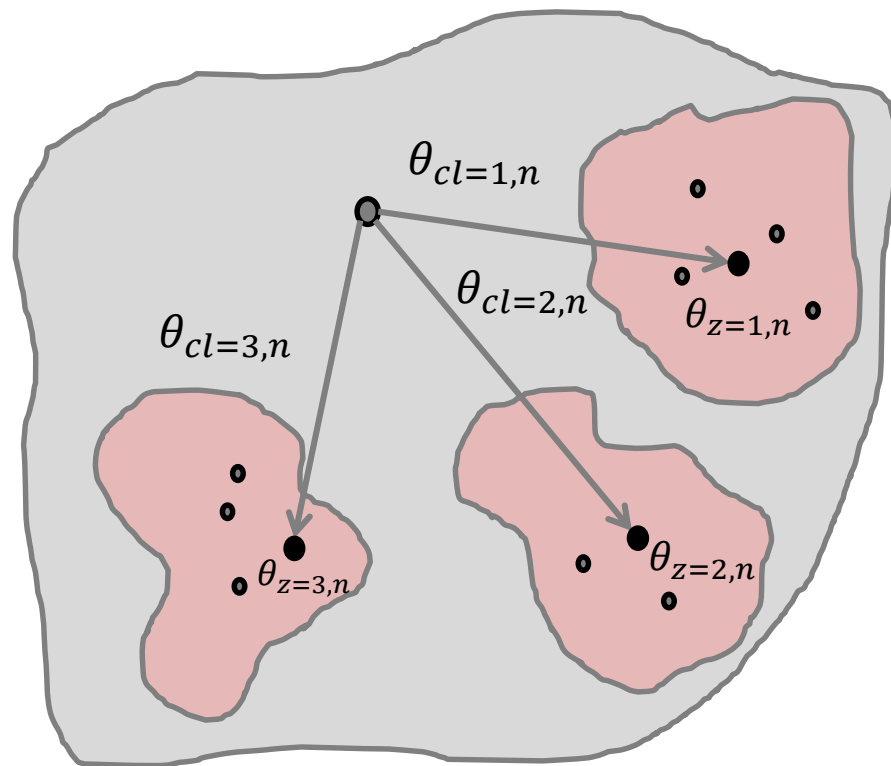
Optimal Clusters





Use of all observations from all clusters

Global Interpolation



Selection of one station in each cluster

Local Interpolation with selected neighbors

Minimize $\|E \cdot x - F\|_2^2$

Subject to:

$$\xi_{\alpha,cl} \geq 0 \quad \forall cl$$

$$\theta_{cl,n} = \sum_{i=1}^{m_{cl}} \theta_{i,n} \quad \forall cl, \quad \forall n$$

$$\theta_{cl,n} = \theta_{z,n} \quad z \in m_{cl} \quad \forall cl, \quad \forall n$$

$$\theta'_{\alpha,n} = \sum_{cl=1}^{Ncl} \theta_{cl,n} \xi_{\alpha,cl} \quad \forall n$$

E is the $no \times cl$ matrix of $\theta_{cl,n}$ values, x is the matrix $cl \times 1$ of $\xi_{\alpha,cl}$ weight values and F is the matrix of $no \times 1$ values of data at base station ($\theta_{\alpha,n}$). Minimization of Square Norm.

- Sum of observations in each cluster, m_{cl} : number of stations in a cluster, cl .

- $\theta_{z,n}$ is site in a cluster with a maximum correlation with base station.

- Missing data estimation using Weights $\xi_{\alpha,cl}$

Minimize $\sum_{n=1}^{nt} |\theta'_{\alpha,n} - \theta_{\alpha,n}|$

Subject to:

$$\theta'_{\alpha,n} = \frac{\sum_{j=1}^{ns-1} \theta_{j,n} \varpi_{\alpha,j}}{\sum_{j=1}^{ns-1} \varpi_{\alpha,j}} \quad \forall n$$

$$\sum_{j=1}^{ns-1} \varpi_{\alpha,j} \leq ng \quad \forall n$$

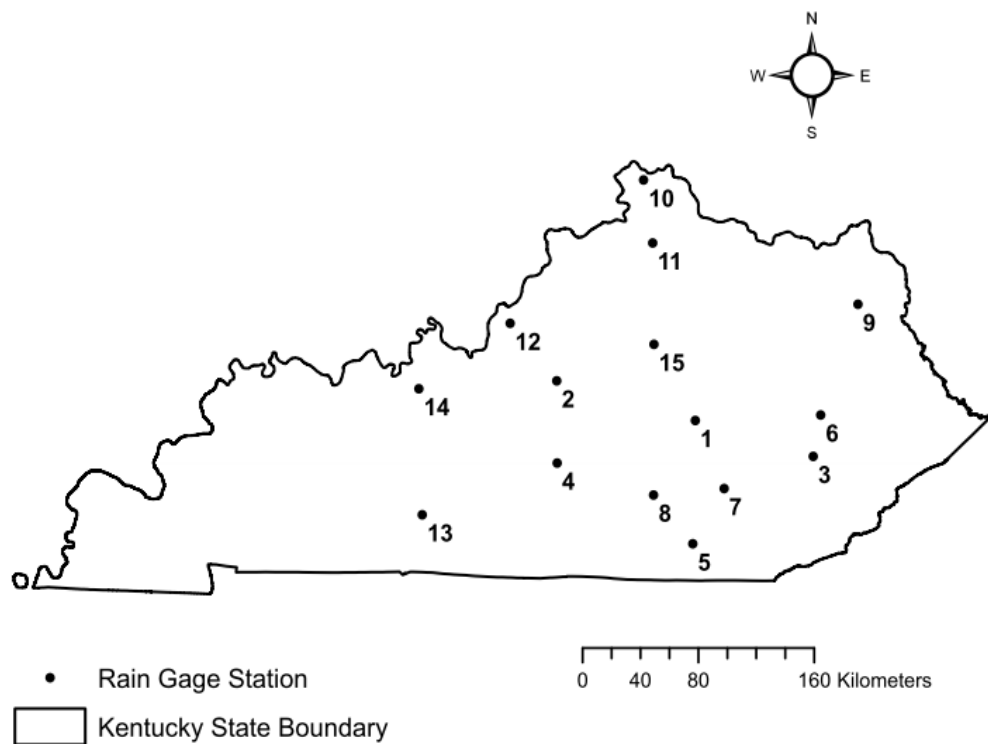
- Optimal mean imputation is possible by selecting optimal set of neighbors
- Binary variable $\varpi_{\alpha,j}$ is used for selection
- ng : number of sites
- When $ng = 1$, the problem is single best estimator (not based on Euclidean distance or correlation criterion)

- Correlation
 - $w_{\rho} = 1 - [\frac{\Omega_{\rho}}{\rho'}(\rho)]$
- Absolute Error
 - $w_{AE} = \frac{\Omega_{AE}}{AE'}(AE)$
- RMSE
 - $w_{RMSE} = \frac{\Omega_{RMSE}}{RMSE'}(RMSE)$
- MAE
 - $w_{MAE} = \frac{\Omega_{MAE}}{MAE'}(MAE)$
- Weighted Error Measure
 - $W_{total} = w_{\rho} + w_{AE} + w_{RMSE} + w_{MAE}$

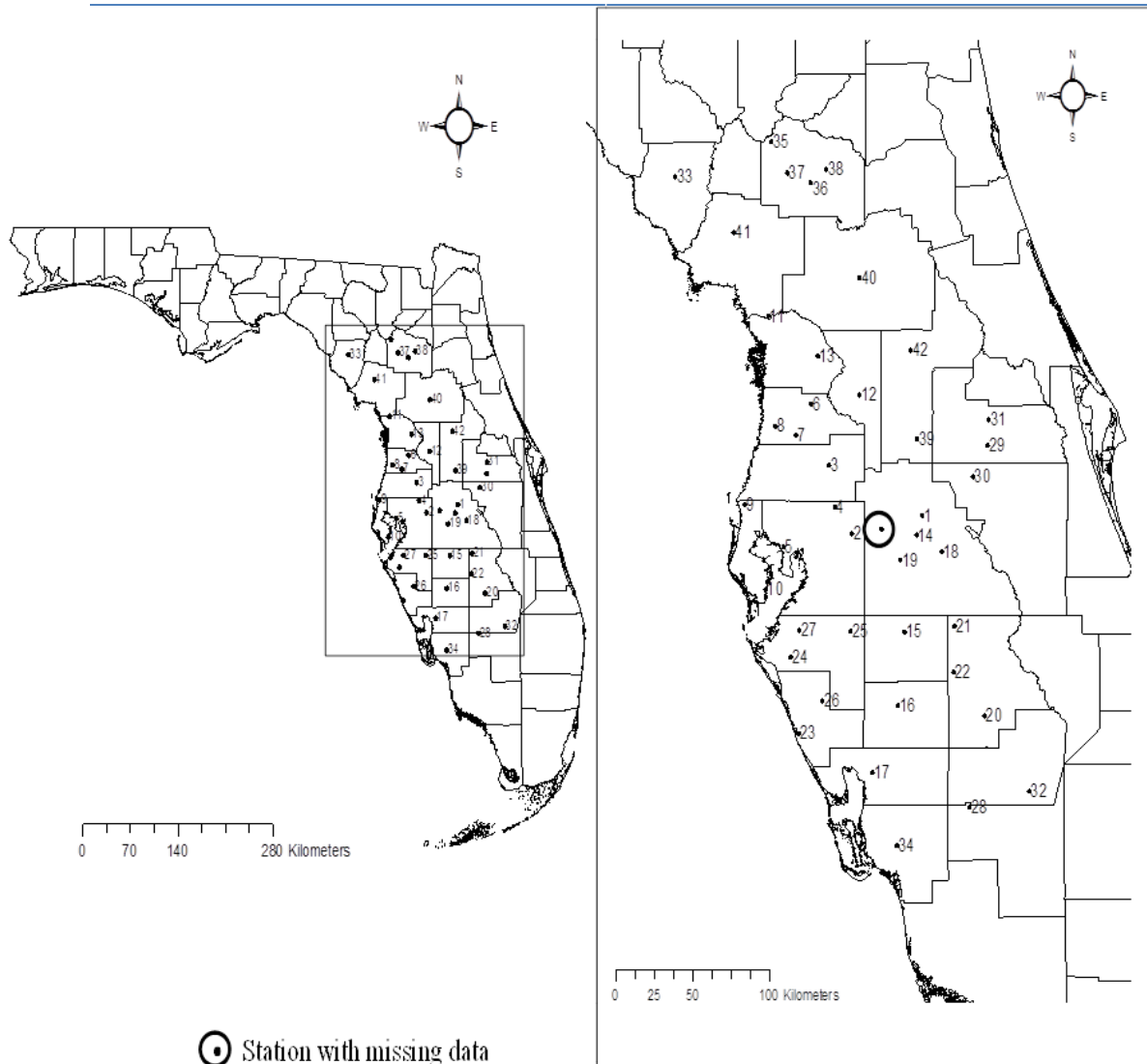
Performance measures are used to evaluate the missing data methods.

The measures are scaled so that they can be added to obtain one single measure to facilitate comparison

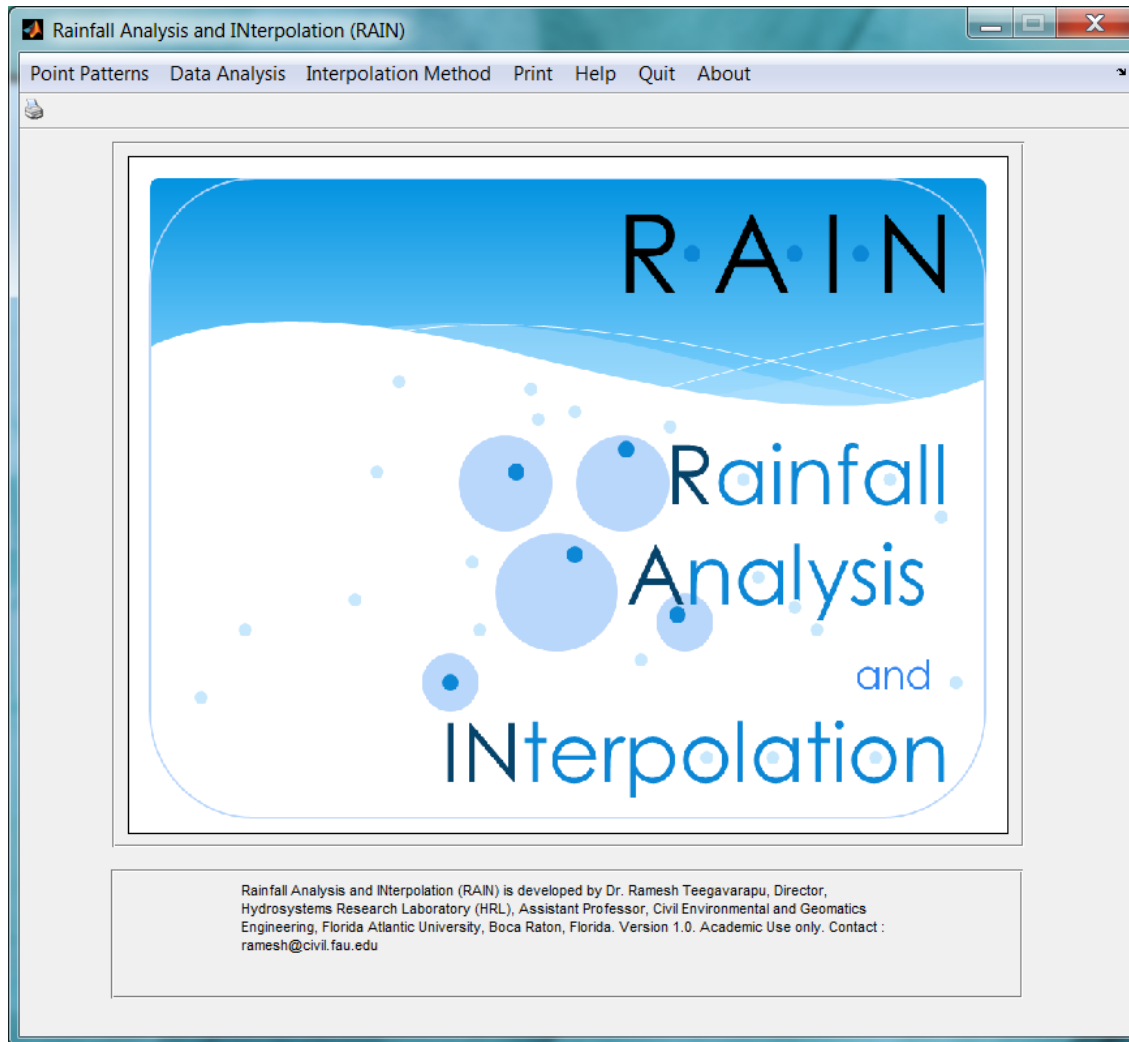
Evaluation is carried out using other visual measures also.



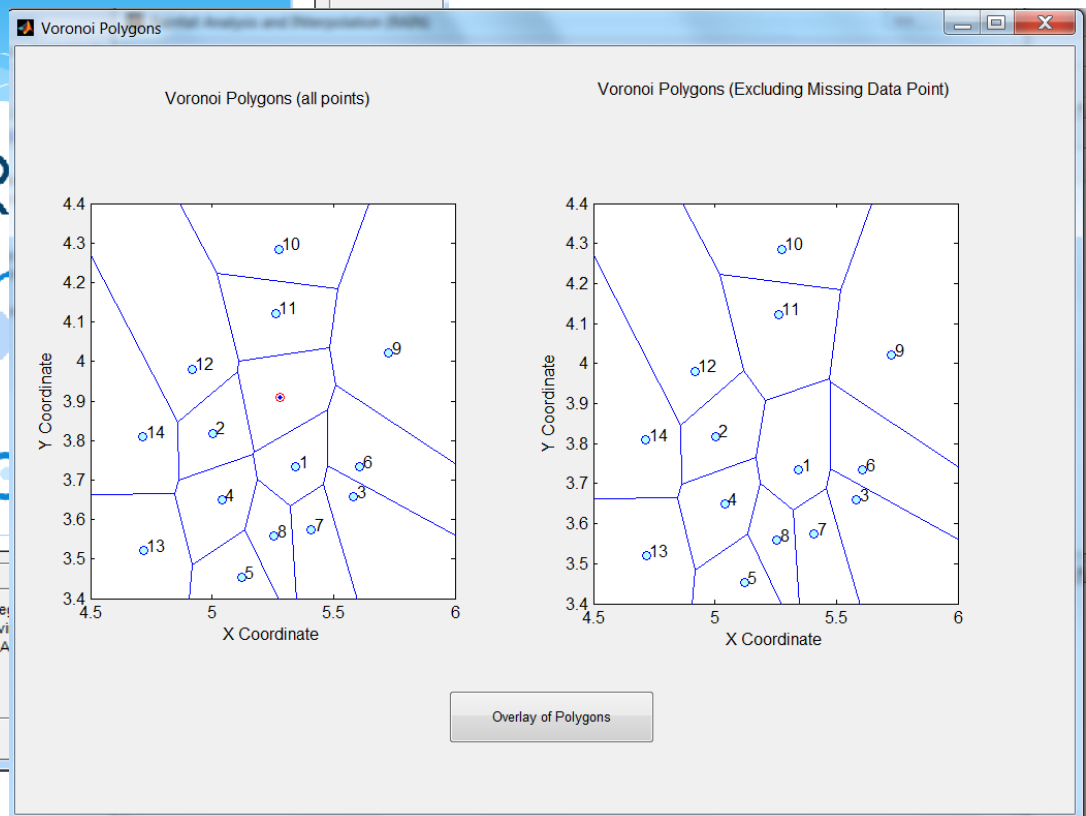
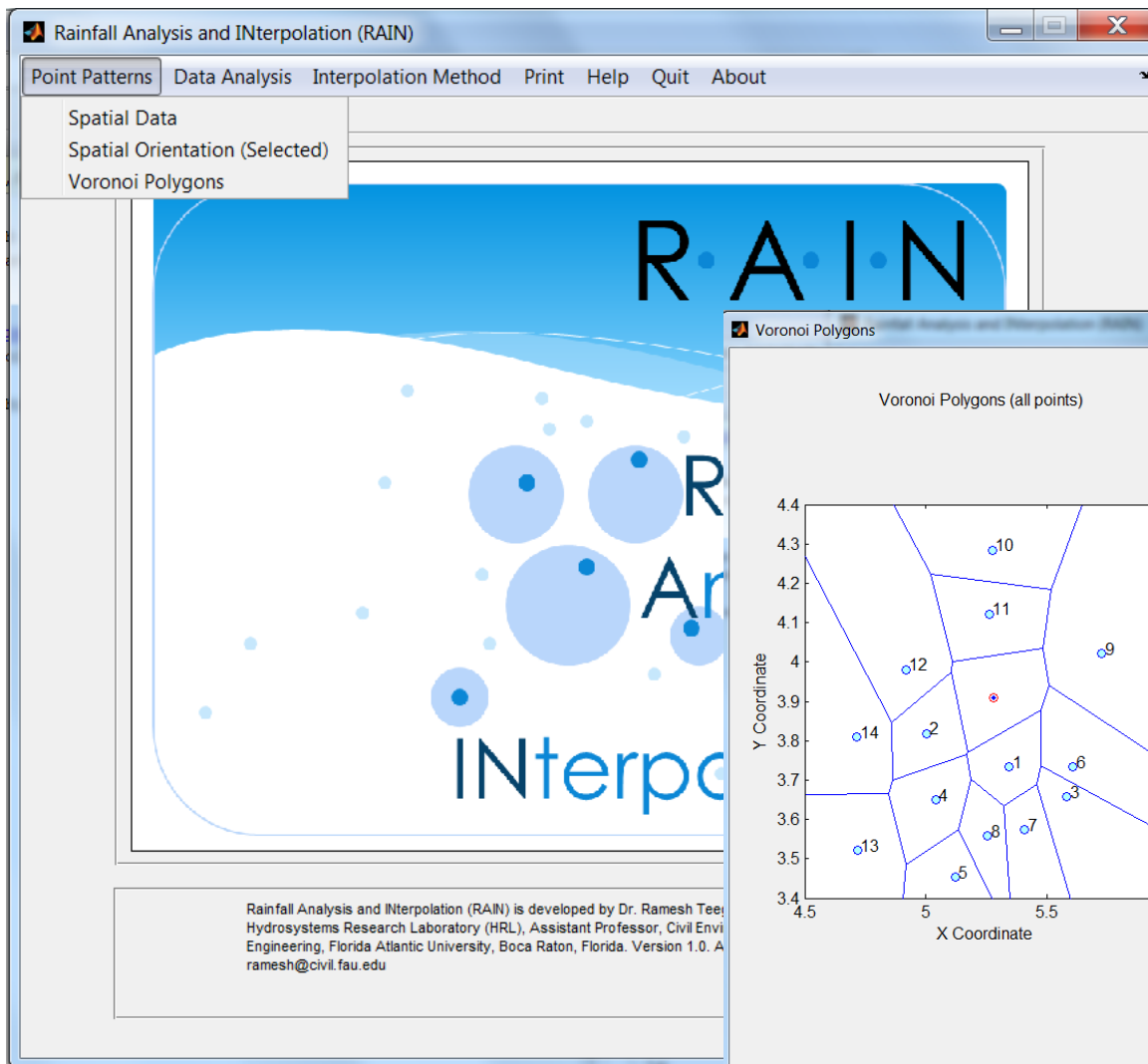
- 15 stations in the state of Kentucky
- Long-term daily precipitation data
- 1971 -2002
- Approximately 70% data used for model development and 30% for testing
- K-fold cross validation

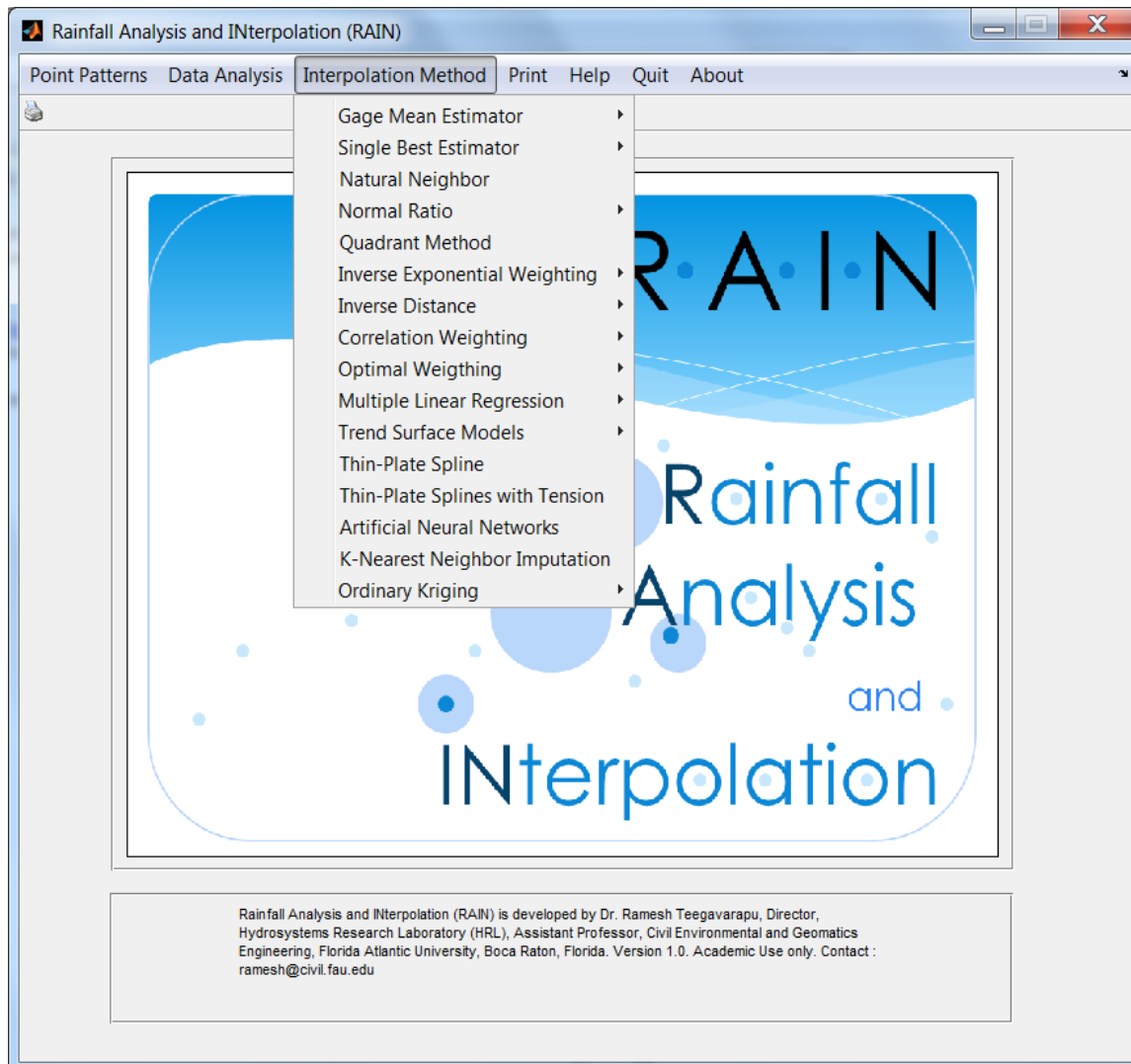


- 43 stations in the state of Kentucky
- Daily precipitation data
- 1994-1999
- Approximately 70% data used for model development and 30% for testing
- K-fold cross validation



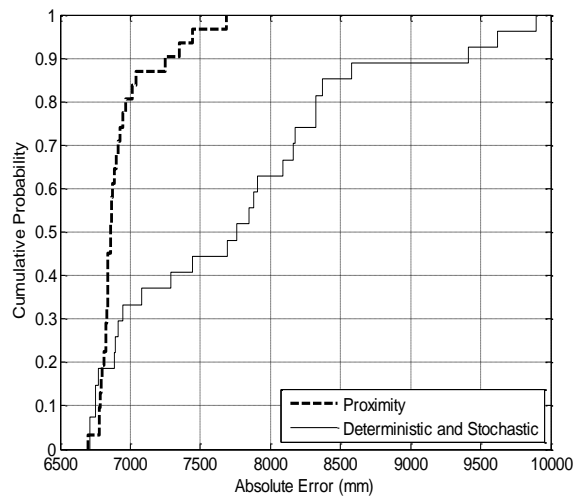
- A suite of deterministic and stochastic interpolation methods
- Inputs: Precipitation data at a location along with spatial coordinates of the observation stations in space.
- Outputs : Correlation matrices, precipitation data , estimated missing precipitation data, scatter graphs, residual plots, semi-variogram plots (kriging), Thiessen polygons, Negative value estimates, optimal parameters
- Jackknife Cross validation feature available.



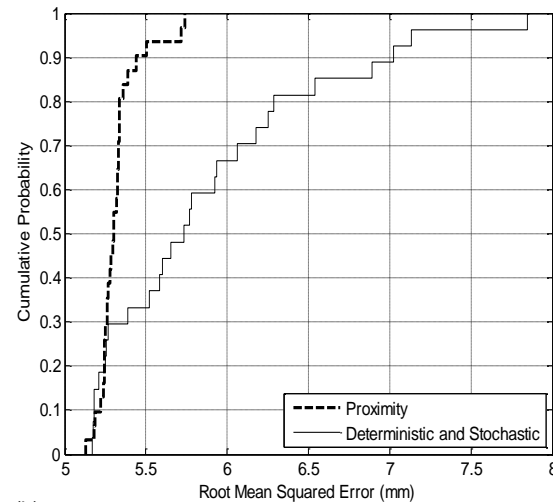


Several different variants
(over 40) of deterministic
and stochastic
Interpolation
Methods

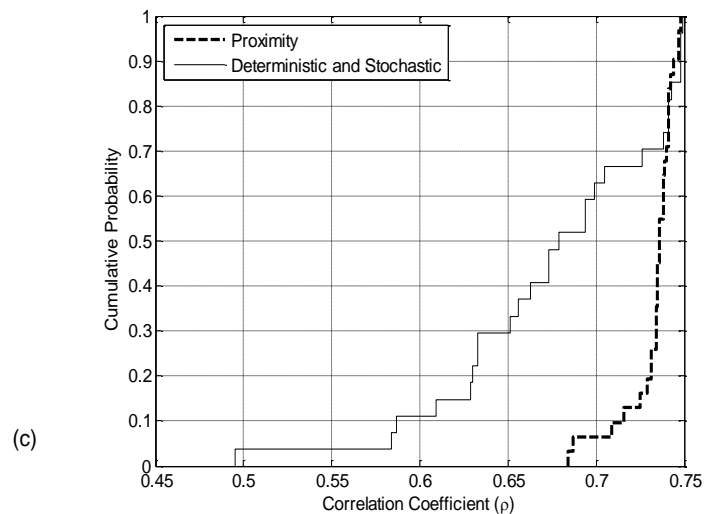
- Deterministic methods
Including new methods
- Ordinary Kriging
 - Gaussian
 - Circular
 - Exponential
 - Spherical



(a)



(b)



(c)

Grouped into two sets: 1)
all optimal proximity-
based and k-nn
classification and k-means
clustering methods

2) Deterministic and
stochastic interpolation
methods for comparison
purposes.

Performance Evaluation (Region I)

Proximity Measure	AE (mm)	MAE (mm)	RMSE (mm)	ρ
Euclidean	6859	2.373	5.335	0.735
Squared Euclidean	6867	2.377	5.318	0.736
Manhattan	6859	2.374	5.334	0.735
Maximum	7348	2.543	5.439	0.716
Minkowski ($\lambda=3$)	6883	2.383	5.284	0.738
Gower	7017	2.428	5.336	0.731
Cosine	6914	2.239	5.250	0.741
Canberra	6899	2.388	5.265	0.739
Correlation	6926	2.397	5.361	0.734
Mahalanobis	7684	2.659	5.715	0.684

The top two methods based on total weights before and after optimization are Minkowski and correlation, and Cosine and Canberra respectively.

The two best methods with and without optimization are Kulzinksky and Jaccard and Yule and Sokal-Michener respectively

Proximity Measure	AE (mm)	MAE (mm)	RMSE (mm)	ρ
Simple Matching	6835	2.366	5.330	0.735
Jaccard	6794	2.352	5.261	0.740
Russell and Rao	6836	2.366	5.264	0.738
Dice	6798	2.353	5.296	0.736
Rogers and Tanimoto	6840	2.368	5.333	0.734
Pearson	6825	2.362	5.323	0.731
Yule	6826	2.363	5.237	0.742
Sokal-Michener	6841	2.367	5.243	0.741
Kulzinksky	6776	2.341	5.279	0.738
Hamming	6837	2.367	5.242	0.741

Region II

Proximity Measure	AE (mm)	MAE (mm)	RMSE (mm)	ρ
Euclidean	2223	3.131	6.841	0.704
Squared Euclidean	2261	3.183	6.994	0.685
Manhattan	2272	3.200	6.722	0.714
Maximum	2420	3.408	7.203	0.660
Minkowski ($\lambda = 3$)	2315	3.259	7.158	0.678
Gower	2446	3.445	6.947	0.692
Cosine	2252	3.176	6.945	0.690
Canberra	2258	3.181	6.831	0.702
Correlation	2256	3.177	7.022	0.684
Mahalanobis	2693	3.793	7.607	0.611

The two best methods before and after optimization are Manhattan and Euclidean and Cosine and squared Euclidean respectively

The two best methods before and after optimization are Sokal-Michener and Russell and Rao, and Pearson and Yule respectively

Proximity Measure	AE (mm)	MAE (mm)	RMSE (mm)	ρ
Simple Matching	2334	3.286	7.171	0.667
Jaccard	2434	3.428	6.927	0.693
Russell and Rao	2356	3.318	6.888	0.696
Dice	2327	3.277	7.143	0.669
Rogers and Tanimoto	2448	3.442	6.967	0.688
Pearson	2423	3.419	6.925	0.693
Yule	2450	3.451	6.961	0.689
Sokal-Michener	2245	3.162	6.804	0.705
Kulzinsky	2311	3.256	7.183	0.666
Hamming	2478	3.489	6.991	0.685

Region I

θ_{th} (mm)	AE (mm)	MAE (mm)	RMSE (mm)	ρ
13	6878	2.381	5.220	0.743
26	6923	2.396	5.238	0.740
39	7054	2.441	5.309	0.733
52	7204	2.516	5.444	0.715
65	7417	2.566	5.508	0.709
78	7950	2.736	5.781	0.671

Region II

θ_{th} (mm)	AE (mm)	MAE (mm)	RMSE (mm)	ρ
13	2252	3.176	6.826	0.703
26	2279	3.208	6.994	0.685
39	2250	3.170	6.657	0.720
52	2471	3.479	8.024	0.595
65	2291	3.228	7.088	0.681
78	2596	3.657	7.533	0.623

- The threshold value (θ_{th}) has a significant influence on the performance of binary proximity measures .
- A binary distance metric defined by Sokal-Michener that performed well in region I and II was used to evaluate this influence on imputation.

Percentage of Data	AE (mm)	MAE (mm)	RMSE (mm)	ρ
10	2409	3.393	7.630	0.621
20	2415	3.453	7.884	0.605
50	2387	3.361	7.528	0.635
70	2342	3.302	7.278	0.657
80	2344	3.302	7.146	0.669

How many Classes ?

Region I Correlation

Number of classes	AE (mm)	MAE (mm)	RMSE (mm)	ρ
14	7039	2.436	5.247	0.741
10	6945	2.404	5.300	0.736
7	7248	2.501	5.327	0.729
3	7445	2.577	5.501	0.709

Optimal weights

Number of classes	AE (mm)	MAE (mm)	RMSE (mm)	ρ
14	6871	2.388	5.221	0.744
10	6811	2.357	5.186	0.747
7	6787	2.349	5.184	0.747
3	6697	2.319	5.128	0.748

Region II

Number of classes	AE (mm)	MAE (mm)	RMSE (mm)	ρ
14	2474	3.484	7.041	0.696
10	2557	3.604	7.608	0.628
7	2448	3.449	7.450	0.658
3	2306	3.248	6.812	0.708

The optimal weighting scheme provided better results compared to correlation weighting in both the regions.

In case of region I, best performance was obtained based on the total weighted score, when ten and three classes are used for correlation and optimization-based methods respectively.

For region II, the best performance was achieved when fourteen and three classes are used for correlation and optimization-based methods respectively.

The methods developed are compared with several deterministic and stochastic spatial interpolation methods

1. Gauge Mean Estimator (distance)
2. Gauge Mean Estimator (correlation)
3. Single Best Estimator (distance)
4. Single Best Estimator (correlation)
5. Normal Ratio(distance)
6. Normal Ratio (correlation)
7. Quadrant (one neighbor)
8. Inverse exponential (radius limited)
9. Inverse Distance (optimal exponent)
10. Inverse Distance
11. Correlation Coefficient Weighting
12. Correlation Coefficient Weighting (optimum exponent)
13. Optimal Weighting (positive weights - all neighbors)
14. Optimal Weighting (positive weights, nearest neighbors-correlation -based)
15. Multiple Linear Regression
16. Step-Wise Regression
17. Robust-fit Regression
18. Trend surface model (global, linear)
19. Trend surface model (global, quadratic)
20. Trend surface model (global, cubic)
21. Thin plate spline
22. Thin plate splines with tension
23. Artificial neural networks
24. Ordinary Kriging (Spherical, Gaussian, Exponential, Circular)

Method	Rank	Method	Rank
K-NN Classification (optimization, 3 classes)	1	Correlation	31
Step-Wise Regression	2	Artificial neural networks	32
Multiple Linear Regression	3	K-Means Cluster (optimization, 4 clusters)	33
Optimal Weighting (postive weights - all neighbors)	4	Normal Ratio (correlation)	34
Optimal Weighting (postive weights, nearest neighbors-correlation based)	5	Gower	35
Cosine	6	Robust-fit Regression	36
K-NN Classification (optimization, 7 classes)	7	K-NN Classification (correlation, 7 classes)	37
K-NN Classification (optimization, 10 classes)	8	K-Means Cluster (optimization, 2 clusters)	38
Kulzinksky	9	Maximum	39
Jaccard	10	Normal Ratio(distance)	40
Yule	11	K-NN Classification (correlation, 3 classes)	41
Hamming	12	Gauge Mean Estimator (distance)	42
Sokal-Michener	13	Trend surface model (global, linear)	43
K-NN Classification (optimization, 14 classes)	14	Mahalanobis	44
K-Means Cluster (optimization, 6 clusters)	15	Inverse exponential (radius limited)	45
Dice	16	Reciprocal Variance Weighting (Spherical Semi-variogram)	46
Russell and Rao	17	Inverse Distance	47
Correlation Coefficient Weighting (optimum exponent)	18	Inverse Distance (optimal exponent)	48
Correlation Coefficient Weighting	19	Thin plate splines with tension	49
Canberra	20	Ordinary Kriging (Guassian Semi-variogram)	50
Simple Matching	21	Natural Neighbor	51
Minkowski ($\lambda = 3$)	22	Ordinary Kriging (Exponential Semi-variogram)	52
Pearson	23	Quadrant (one neighbor)	53
Rogers and Tanimoto	24	Ordinary Kriging (Circular Semi-variogram)	54
Squared Euclidean	25	Ordinary Kriging (Spherical Semi-variogram)	55
Euclidean	26	Thin plate spline	56
Manhattan	27	Single Best Estimator (correlation)	57
Gauge Mean Estimator (correlation)	28	Trend surface model (global, quadratic)	58
K-NN Classification (correlation, 10 classes)	29	Trend surface model (global, cubic)	59
K-NN Classification (correlation, 14 classes)	30	Single Best Estimator (distance)	60

Method	Rank	Method	Rank
Normal Ratio (correlation)	1	Pearson	31
Ordinary Kriging (Circular Semi-Variogram)	2	Dice	32
Thin plate splines with tension	3	Kulzinsky	33
Robust-fit Regression	4	K-NN Classification (correlation, 14 classes)	34
Euclidean	5	Jaccard	35
Manhattan	6	Step-Wise Regression	36
Sokal-Michener	7	Reciprocal Variance Weighting (Spherical Semi-variogram)	37
Inverse Distance (optimal exponent)	8	Simple Matching	38
Correlation Coefficient Weighting (optimal exponent)	9	Gower	39
Artificial neural networks	10	Rogers and Tanimoto	40
Gauge Mean Estimator (correlation)	11	Yule	41
Optimal Weighting (postive weights, nearest neighbors-correlation based)	12	K-NN Classification (optimization, 14 classes)	42
K-NN Classification (optimization, 3 classes)	13	Hamming	43
Optimal Weighting (postive weights - all neighbors)	14	Maximum	44
Cosine	15	Multiple Linear Regression	45
Quadrant (one neighbor)	16	Trend surface model (global, cubic)	46
Inverse Distance	17	Trend surface model (global, quadratic)	47
Squared Euclidean	18	Single Best Estimator (distance)	48
Ordinary Kriging (Spherical Semi-variogram)	19	K-NN Classification (correlation, 7 classes)	49
Inverse exponential (radius limited)	20	K-NN Classification (optimization, 7 classes)	50
Natural Neighbor	21	K-NN Classification (correlation, 3 classes)	51
Correlation	22	K-NN Classification (correlation, 10 classes)	52
Russell and Rao	23	Trend surface model (global, linear)	53
Correlation Coefficient Weighting	24	K-Means Cluster (optimization, 2 clusters)	54
Gauge Mean Estimator (distance)	25	K-NN Classification (optimization, 10 classes)	55
K-Means Cluster (optimization, 6 clusters)	26	Ordinary Kriging (Exponential Semi-variogram)	56
Thin plate spline	27	Mahalanobis	57
Normal Ratio(distance)	28	Artificial neural networks	58
K-Means Cluster (optimization, 4 clusters)	29	Single Best Estimator (correlation)	59
Minkowski ($\lambda = 3$)	30	Ordinary Kriging (Guassian Semi-variogram)	60

-
- Proximity metrics based principles of numerical taxonomy are used for development of new alternatives to distance based spatial interpolation.
 - Ten real and binary metrics are used to evaluate optimal spatial interpolation methods for estimation of missing precipitation data at a site. Several others can be evaluated.
 - Different threshold values for binary transformations are evaluated and the results seem to be sensitive to the threshold values with lower values improving the performance of the spatial interpolation methods.
 - A total of twenty proximity metrics are evaluated and all most all the metrics seem to perform better than several traditional spatial interpolation methods.
 - Temporal interpolation is not possible due low serial autocorrelation at a temporal scale of one day.

-
- Cluster-based spatial interpolation methods are evaluated in this study
 - Clusters are identified by **k-means clustering** approach.
 - Best performance measures were obtained when one single station in each station with highest correlation was used in the weighting scheme.
 - Optimal number of spatial clusters can be found using this methodology.
 - Classifier-based spatial interpolation methods are evaluated in this study
 - Classification is achieved using **K-nearest neighbor technique**.
 - Different distance measures for proximity were evaluated .
 - Number of classes has an effect on the performance measures of the method.
 - Variants of traditional spatial interpolation methods using optimal mean imputation, optimal selection of neighbors and other minor variants of traditional methods are developed for comparison.

Limitations of Deterministic Interpolation Methods

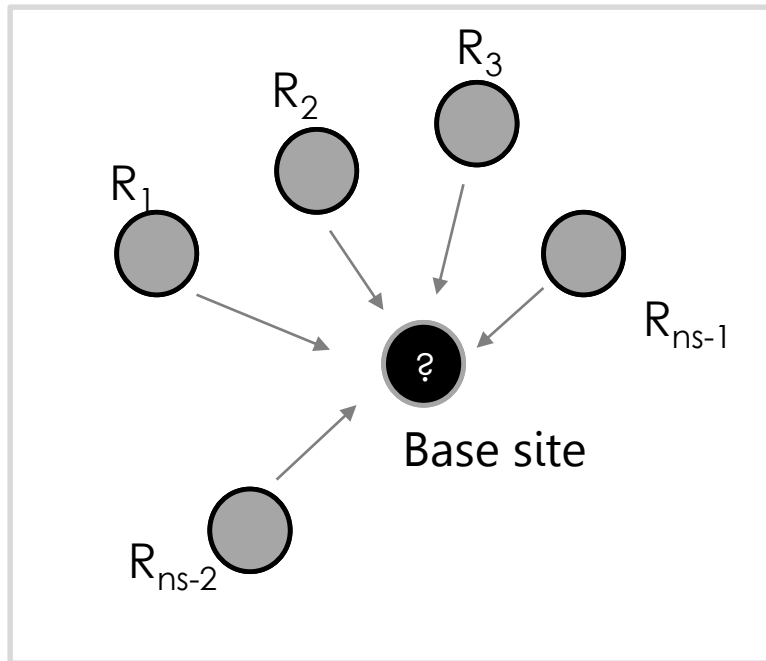
- Single imputation (single estimate).
- Lack of uncertainty assessment.
- Lack of mechanism to address variability in data and parameters of the interpolation method.
- Rigid structure of interpolation method with no consideration of parameter uncertainty.
- Local-global interpolation.
- Single site or multi-site.
- Process-based (rainfall generating mechanism) interpolation
- Isotropy (or anisotropy)

Uncertainty Assessment of Spatially Interpolated Estimates

- Measurement Errors
- Spatio-Temporal Precipitation Data length
- Precipitation patterns
- Climatic region
- Topographical features
- Temporal resolution
- Data Representativeness
- Homogeneity (change in station location)
- Data window
- Monitoring network orientation

- Objectives
 - Single site, multi-site, preservation of statistical properties and associations

Estimation of Missing Values at a Site



Estimation of missing data at a single site (base site) using available data all other observation sites

$R_1, R_2, \dots, R_{ns-1}$: Rain gage sites
[excluding base site]

Base site: A site with missing data exists

nf : Number of sites selected.

Multiple Imputation

- Multiple Imputation refers to filling of each missing value by a vector of $D > 2$ imputed values.
- The imputed data sets referred to as “multiply imputed data”
- The filled values are referred to as “imputes”.
- Multiple Imputation (MI) can be possible using
 - Single model
 - Multiple models
 - Multiple information sources (by data collectors and analysts)

Multiple Imputation

- Multiple imputation inference involves three distinct phases:
 - The missing data are filled in m times to generate m complete data sets.
 - The m complete data sets are analyzed using standard statistical analyses.
 - The results from the m complete data sets are combined to produce inferential results.
- The MI procedure creates multiply imputed data sets for incomplete multivariate data. It uses methods that incorporate appropriate variability across the m imputations.

Single Imputation

- Mean Imputation (unconditional)
 - Using observed data with no separation of data into classes.
- Mean Imputation (conditional)
 - Using observed variables when the data is separated into classes (regression).
- Stochastic Regression and others
- Replaces missing values by predicted values via regression plus a residual (drawn to reflect uncertainty in the predicated value) (Little and Rubin, 2005).
- A normally distributed residual term to each predicted value restores variability to the imputed data.
- Other methods in social sciences: Expectation maximization, Maximum likelihood, hot-deck imputation and list-wise deletion and MI.

Neighborhood and data window selection

Space: Neighborhood Points Selection

- The number of points and specific combination of points (rain gage sites) selected will impact the interpolation.

Time: Data Window Selection

- The data window from which the temporal data is used for stochastic and deterministic interpolation used will affect the estimates of missing precipitation data. .

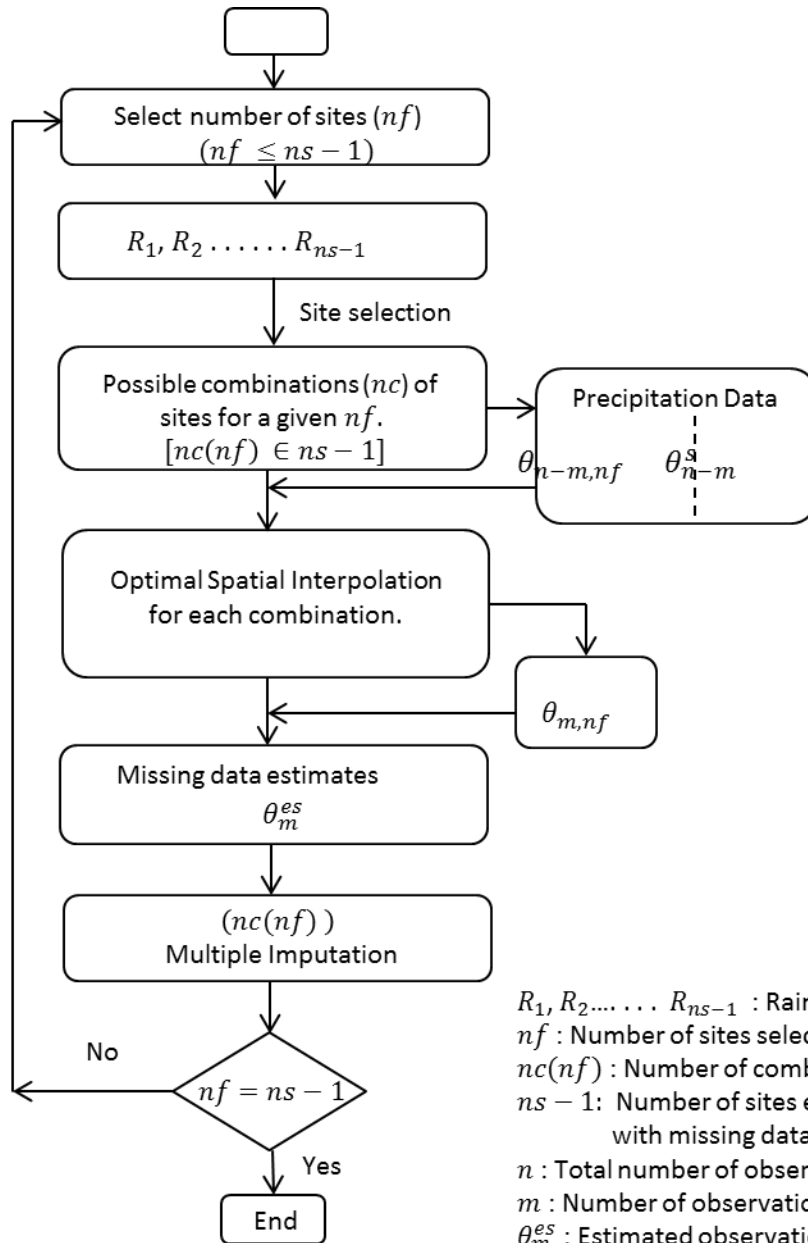


Figure (a)

$R_1, R_2, \dots, R_{ns-1}$: Rain gage sites [excluding base site]
 nf : Number of sites selected.
 $nc(nf)$: Number of combinations for a given nf .
 $ns - 1$: Number of sites excluding site with missing data.
 n : Total number of observations.
 m : Number of observations used as a test set.
 θ_m^{es} : Estimated observations at base site, s .
 $\theta_{n-m,ns-1}$: Rain gage observations.
 N : Maximum number of bootstrap samples

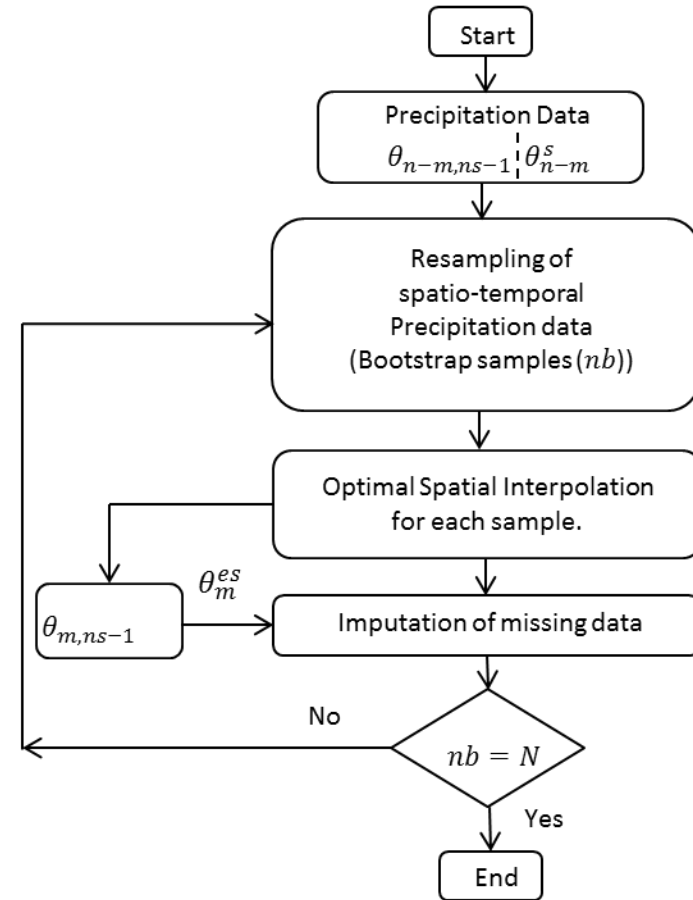
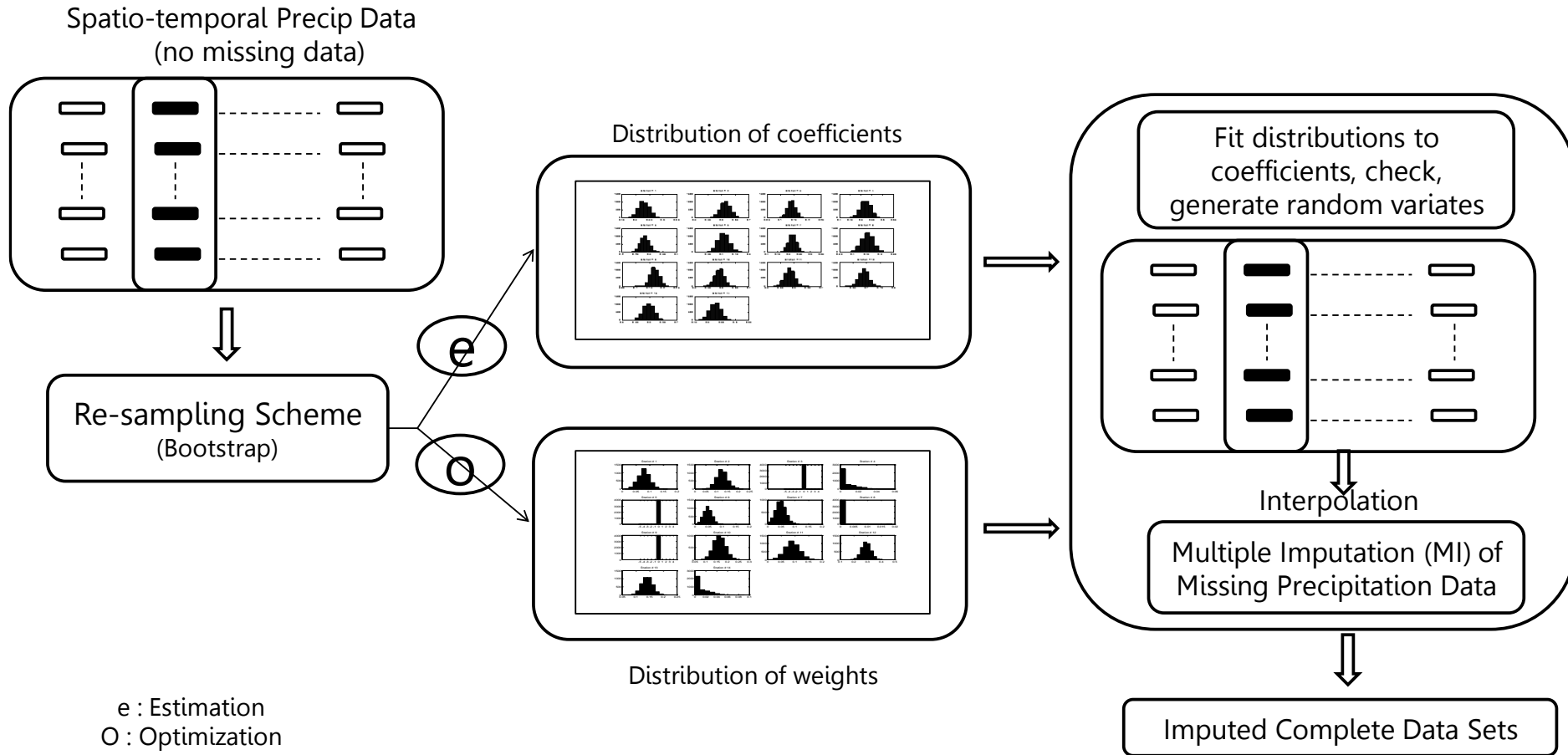


Figure (b)

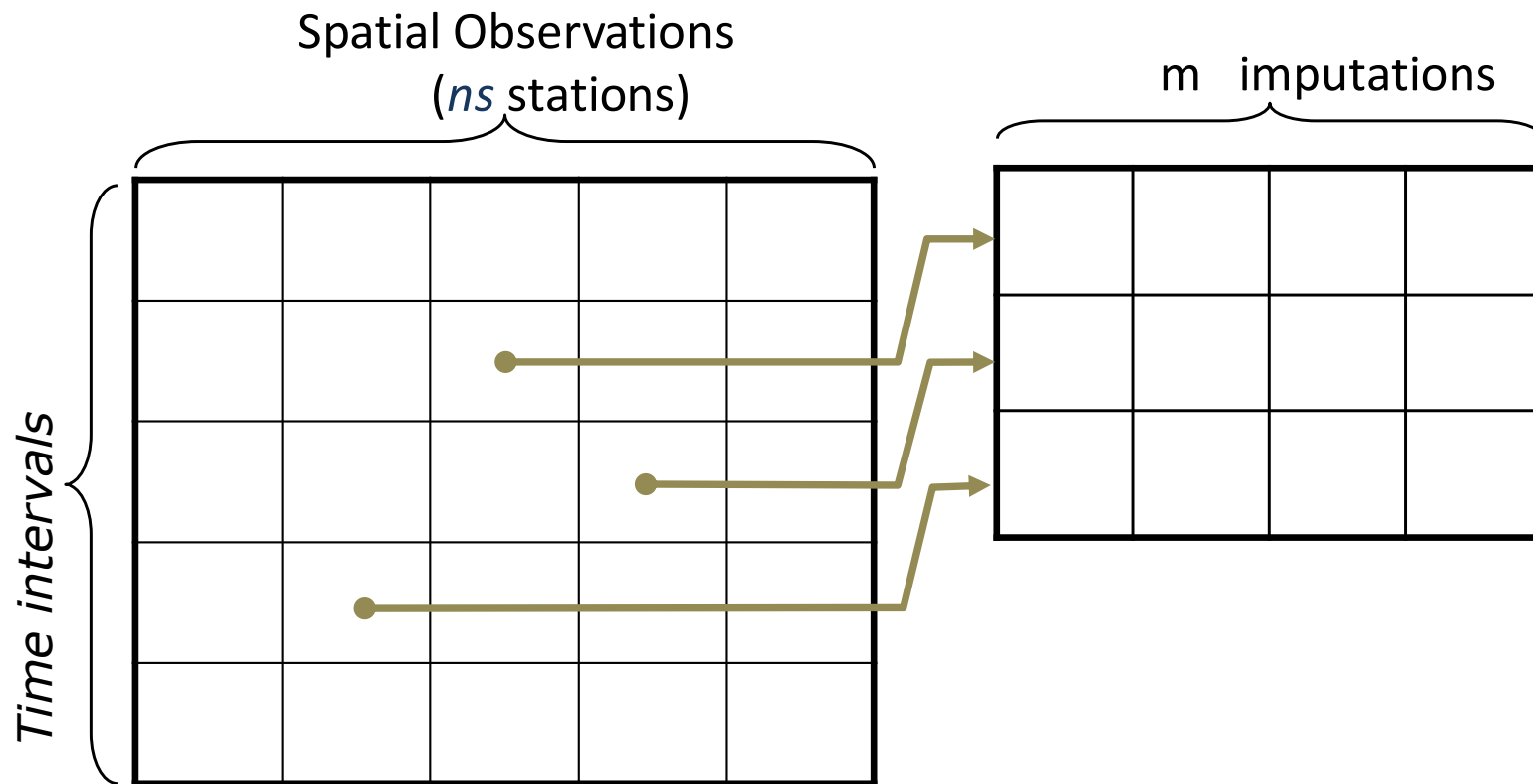
Multiple Imputation



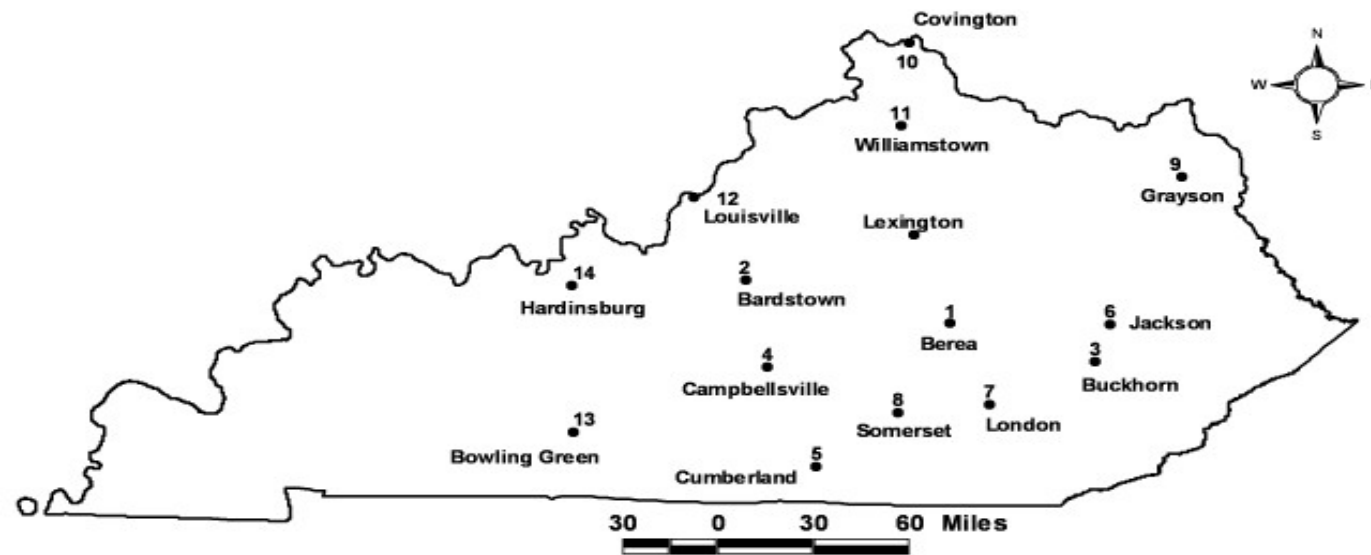
Multiple Imputation

- Multiple imputation using (optimal) spatial interpolation method is possible.
- The spatio-temporal precipitation data (with no missing data) set is used to obtain a number of samples (re-samples) using Bootstrap resampling. Once the samples are obtained optimal spatial interpolation or CCWM is used to obtain the interpolation weights.
- Two ways of multiple imputation are possible:
 - The interpolation weights can be used for development of multiply imputed data sets.
 - Distributions of the weights can be developed to derive random variates of the same to use them in the interpolation.
- Once the imputed data sets are developed, multiple imputations can be used for assessment of uncertainty in precipitation estimates.

Multiple Imputations



Case Study Domain



A network of 15 rain gages in the State of Kentucky is used for the evaluation of the methodology. The base station chosen is Lexington, KY.

Approximately 30 years of daily data are available at all the sites.

-
- The precipitation data at 15 stations is grouped in to 17 equal sized historical data sets in **chronological order**.
 - The minimum and maximum values of correlations provided indicate the wide variation in these values.
 - The average values are close to the values obtained for the entire complete data set comprising of all 17 data sets.
 - The minimum variance is observed for station 3 and maximum for station 6.
 - Skewed distributions of correlation coefficients for different data sets are evident.

Estimates using Correlation weighting and Optimization

Correlation weighting Interpolation

$$\hat{\theta}_m^l(x_m, y_m) = \frac{\sum_{j=1}^{ns-1} w_{mj}^k \theta_j(x_j, y_j)}{\sum_{j=1}^{ns-1} w_{mj}^k} \quad \forall j$$

$$\rho_{mj} = \frac{\sum_{i=1}^{no} (\theta_m^i - \mu_m)(\theta_j^i - \mu_j)}{(no-1)\sigma_m\sigma_j} \quad \forall j$$

$$w_{mj}(x, y) = \rho_{mj} \quad \forall j$$

$\hat{\theta}_m^l$ estimated value of precipitation at the base station m

θ_m^l actual observation in time interval l

θ_j^j ns is the total number of stations including base station
observation at station j

w_{mj}^k weight associated in relation to station j to the station m

Minimize

$$\sqrt{\frac{1}{n} \sum_{l=1}^n (\hat{\theta}_m^l - \theta_m^l)^2}$$

Subject to

$$\hat{\theta}_m^l = \frac{\sum_{j=1}^{ns-1} w_{mj}^k \theta_j^l}{\sum_{j=1}^{ns-1} w_{mj}^k} \quad \forall l$$

$$wl_j \leq w_{mj} \leq wu_j \quad \forall j$$

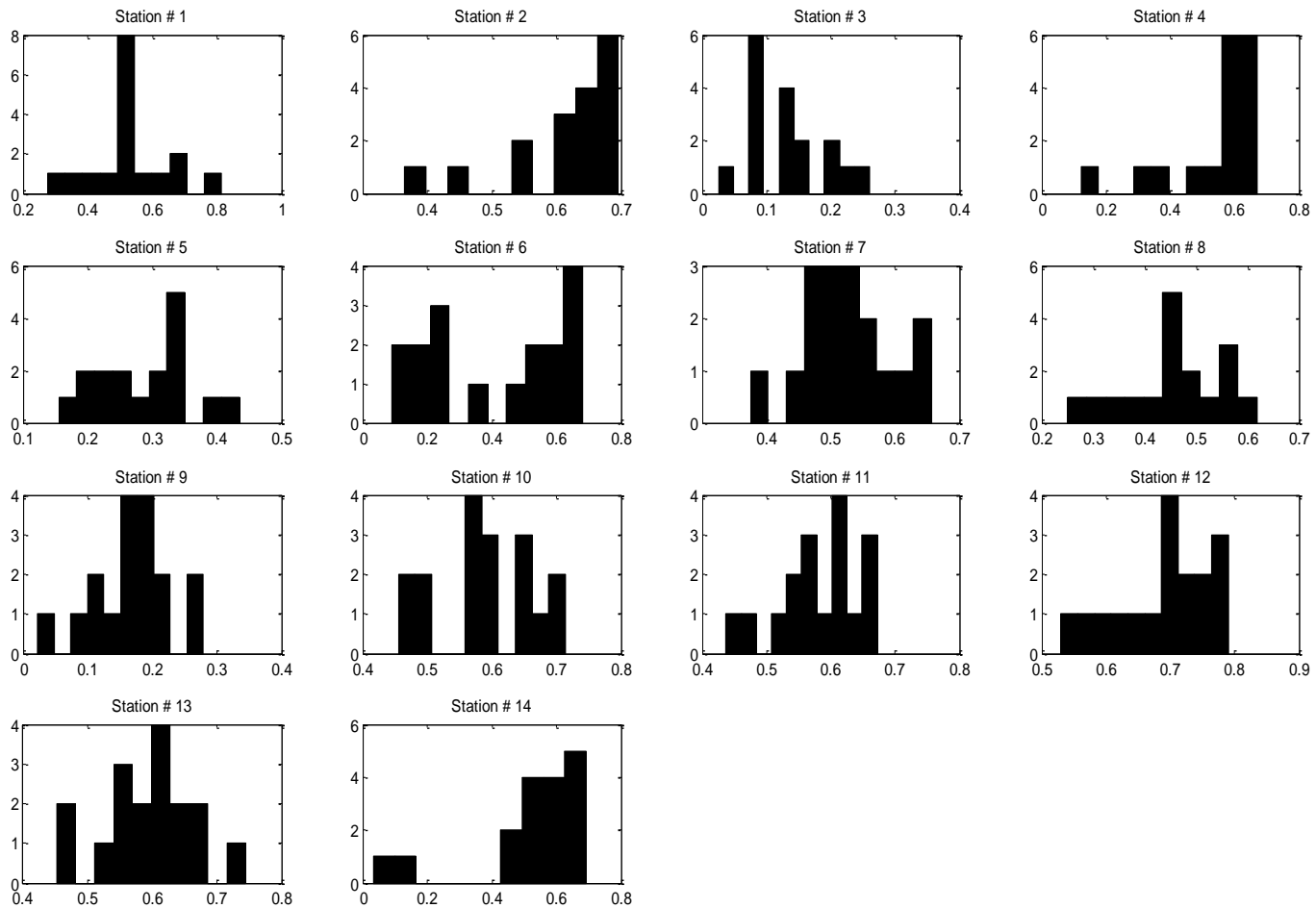
Correlations

Correlation Coefficient														
Station														
Data Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.524	0.635	0.090	0.498	0.191	0.119	0.492	0.438	0.100	0.604	0.614	0.706	0.546	0.034
2	0.621	0.558	0.164	0.624	0.318	0.208	0.644	0.571	0.199	0.488	0.546	0.566	0.622	0.157
3	0.525	0.696	0.261	0.585	0.215	0.269	0.548	0.477	0.181	0.689	0.623	0.710	0.610	0.651
4	0.686	0.676	0.090	0.566	0.287	0.092	0.484	0.396	0.166	0.581	0.546	0.707	0.667	0.691
5	0.815	0.639	0.201	0.564	0.435	0.219	0.543	0.532	0.192	0.585	0.667	0.726	0.531	0.601
6	0.538	0.599	0.140	0.600	0.206	0.182	0.551	0.547	0.204	0.652	0.672	0.761	0.623	0.664
7	0.528	0.600	0.079	0.622	0.337	0.250	0.656	0.566	0.155	0.503	0.568	0.790	0.646	0.548
8	0.276	0.439	0.080	0.327	0.235	0.487	0.496	0.288	0.086	0.457	0.470	0.588	0.597	0.432
9	0.507	0.652	0.025	0.641	0.242	0.508	0.523	0.448	0.195	0.645	0.572	0.735	0.583	0.576
10	0.661	0.667	0.193	0.641	0.336	0.610	0.574	0.494	0.272	0.562	0.528	0.668	0.563	0.599
11	0.377	0.365	0.091	0.121	0.155	0.375	0.375	0.251	0.176	0.455	0.437	0.530	0.470	0.521
12	0.520	0.537	0.131	0.523	0.381	0.627	0.613	0.340	0.111	0.574	0.650	0.783	0.636	0.647
13	0.533	0.681	0.095	0.616	0.343	0.634	0.443	0.441	0.223	0.661	0.632	0.752	0.619	0.675
14	0.455	0.665	0.135	0.372	0.306	0.598	0.463	0.424	0.175	0.578	0.592	0.627	0.543	0.543
15	0.577	0.677	0.141	0.620	0.335	0.656	0.536	0.619	0.150	0.714	0.613	0.778	0.679	0.609
16	0.421	0.659	0.149	0.565	0.264	0.539	0.474	0.439	0.023	0.672	0.610	0.659	0.452	0.453
17	0.537	0.615	0.220	0.671	0.333	0.682	0.513	0.452	0.279	0.589	0.577	0.698	0.745	0.514
Minimum	0.276	0.365	0.025	0.121	0.155	0.092	0.375	0.251	0.023	0.455	0.437	0.530	0.452	0.034
Maximum	0.815	0.696	0.261	0.671	0.435	0.682	0.656	0.619	0.279	0.714	0.672	0.790	0.745	0.691
Average	0.535	0.609	0.134	0.538	0.289	0.415	0.525	0.454	0.170	0.589	0.583	0.693	0.596	0.525
Standard Deviation	0.122	0.090	0.060	0.142	0.074	0.209	0.072	0.099	0.064	0.079	0.064	0.078	0.074	0.179
All Data	0.530	0.614	0.141	0.520	0.283	0.407	0.516	0.452	0.170	0.592	0.584	0.690	0.597	0.527

Observations

- The precipitation data at 15 stations is grouped in to 17 equal sized historical data sets in **chronological order**.
- The minimum and maximum values of correlations provided indicate the wide variation in these values.
- The average values are close to the values obtained for the entire complete data set comprising of all 17 data sets.
- The minimum variance is observed for station 3 and maximum for station 6.
- Skewed distributions of correlation coefficients for different data sets are evident.

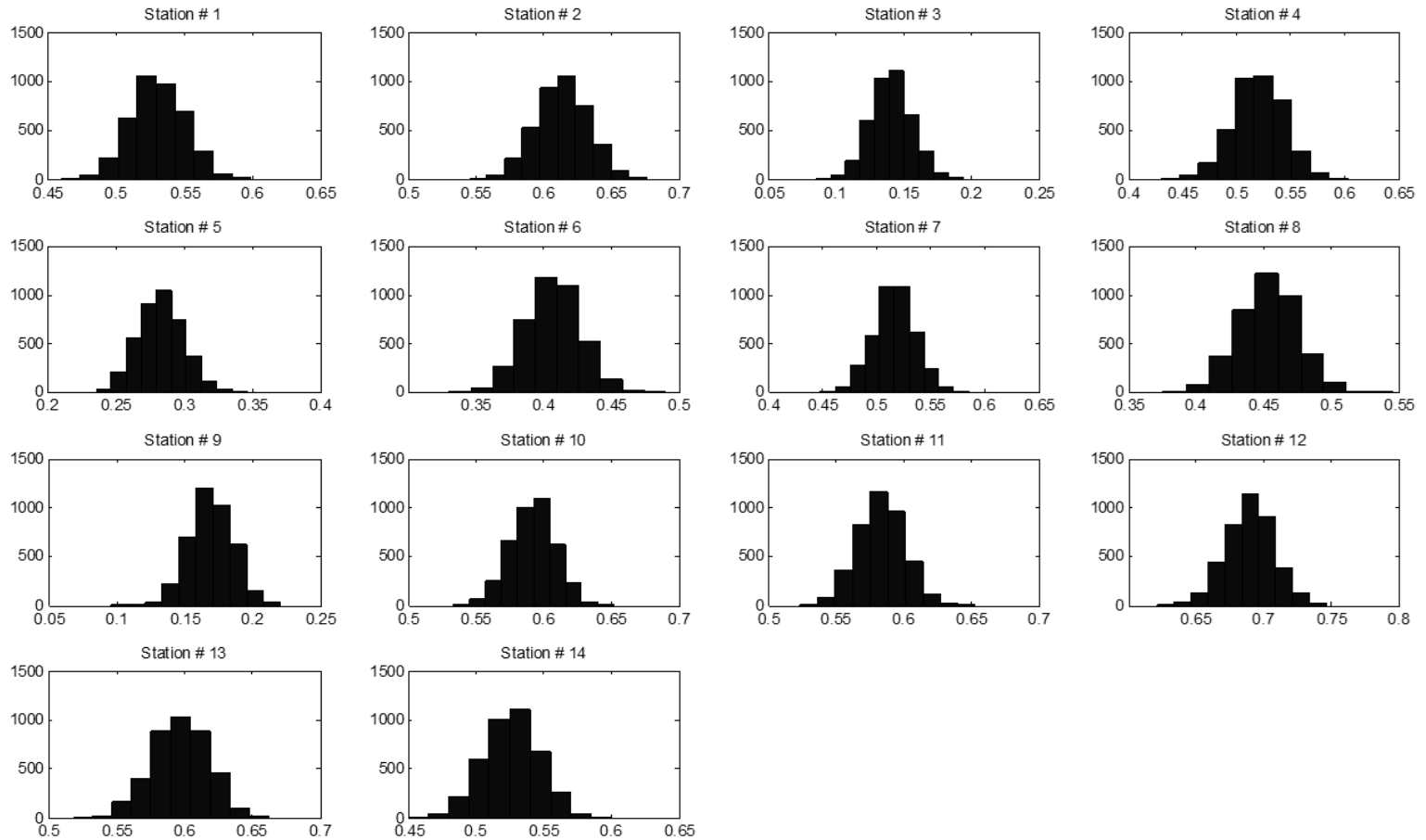
Weights



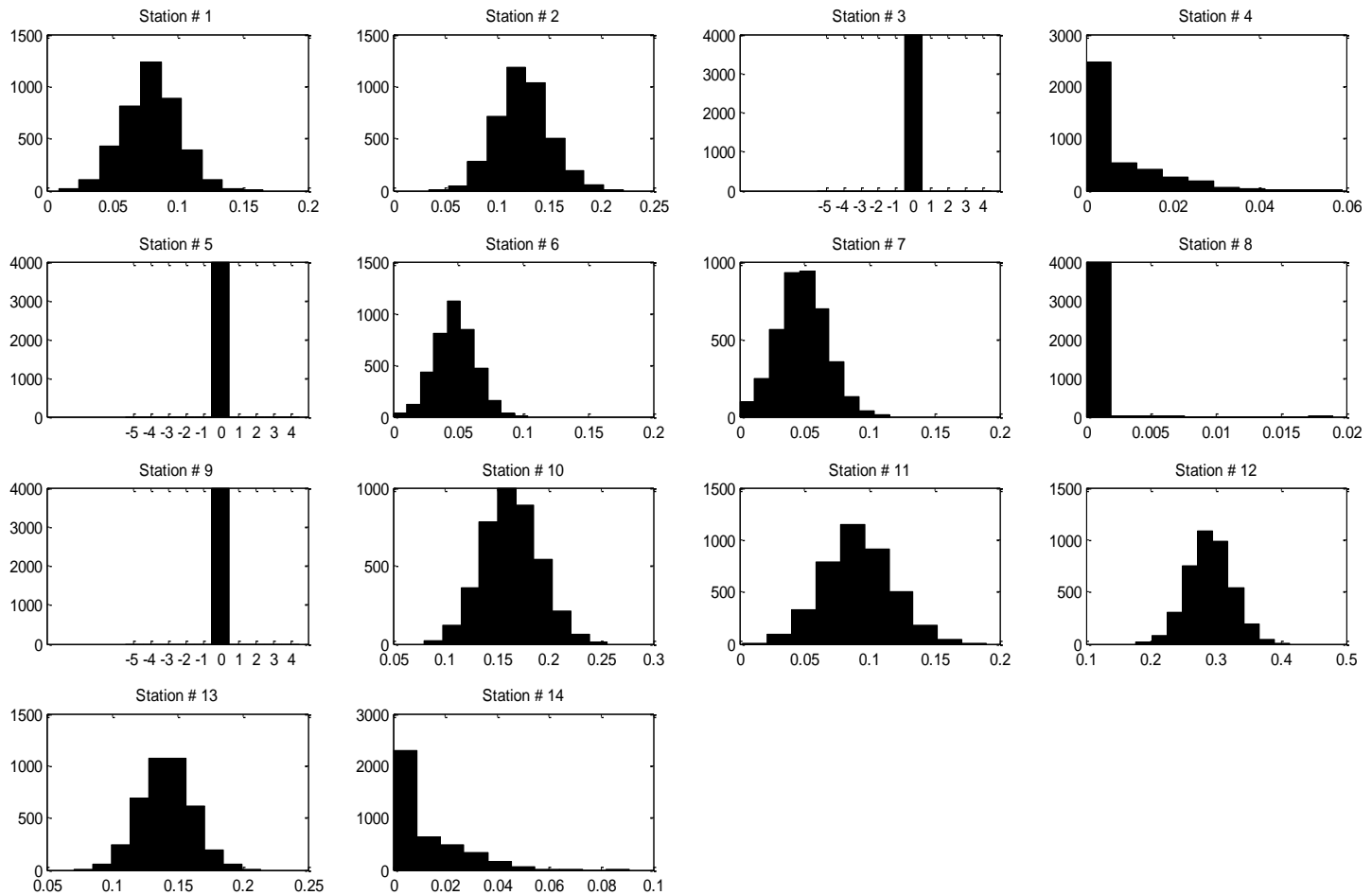
Distributions of Weights

- Distributions of weights at all the sites based on temporally separated historical data sets for model development.
- Highly skewed distributions of weights introduce higher uncertainty in the precipitation estimates

Correlations



Weights



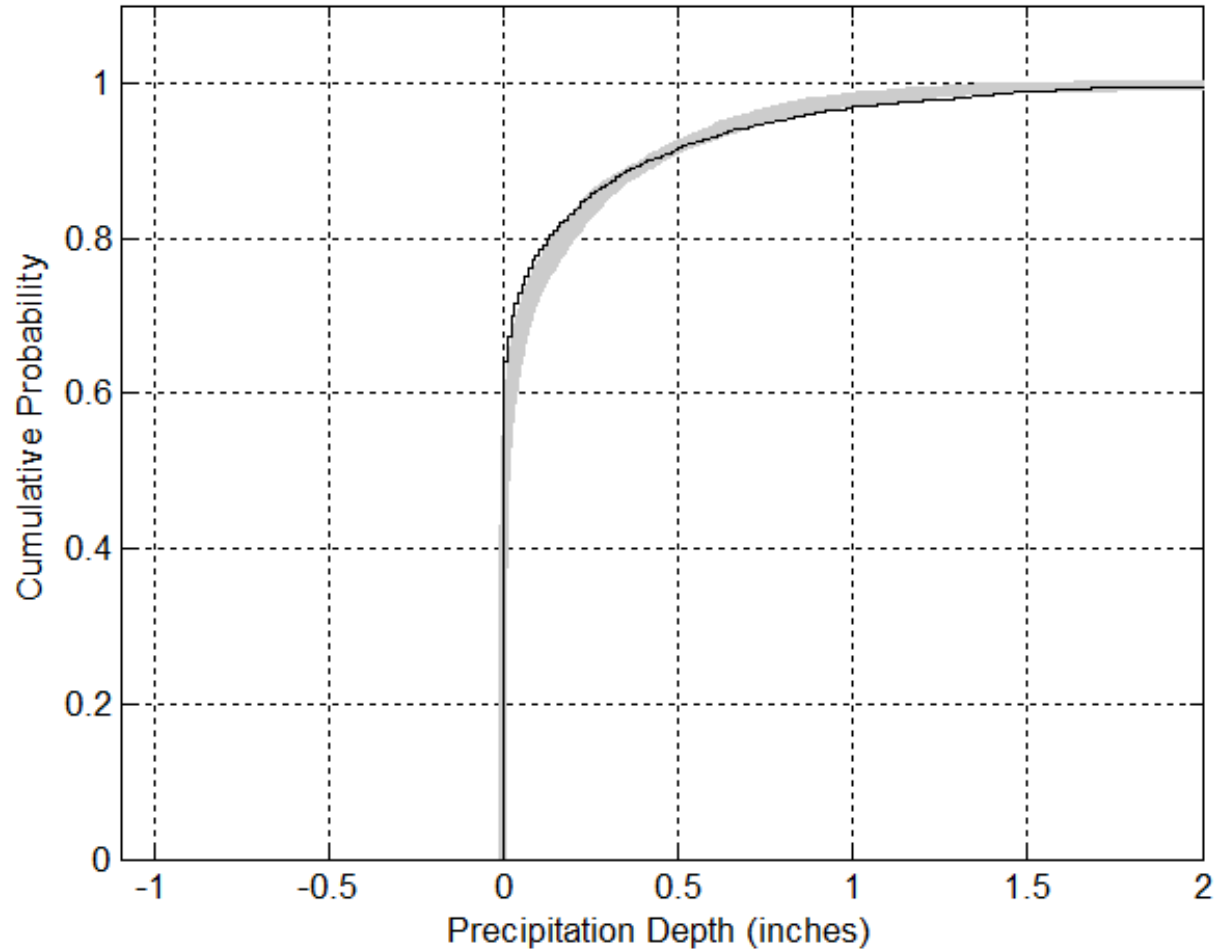
Variability of Performance Measures

	Exponent				
	1	2	4	6	8
Data Set	Correlation Coefficient (ρ)				
1	0.736	0.747	0.757	0.758	0.753
2	0.717	0.712	0.700	0.687	0.674
3	0.735	0.744	0.750	0.752	0.752
4	0.733	0.739	0.741	0.740	0.739
5	0.729	0.735	0.737	0.727	0.706
6	0.734	0.742	0.748	0.749	0.746
7	0.729	0.733	0.737	0.737	0.731
8	0.738	0.744	0.742	0.731	0.717
9	0.737	0.745	0.751	0.752	0.748
10	0.728	0.734	0.739	0.740	0.740
11	0.743	0.750	0.747	0.737	0.728
12	0.736	0.746	0.754	0.752	0.744
13	0.738	0.748	0.756	0.758	0.756
14	0.739	0.751	0.763	0.767	0.767
15	0.733	0.742	0.751	0.754	0.754
16	0.738	0.750	0.759	0.758	0.755
17	0.730	0.736	0.736	0.729	0.719
Average	0.734	0.741	0.745	0.743	0.737

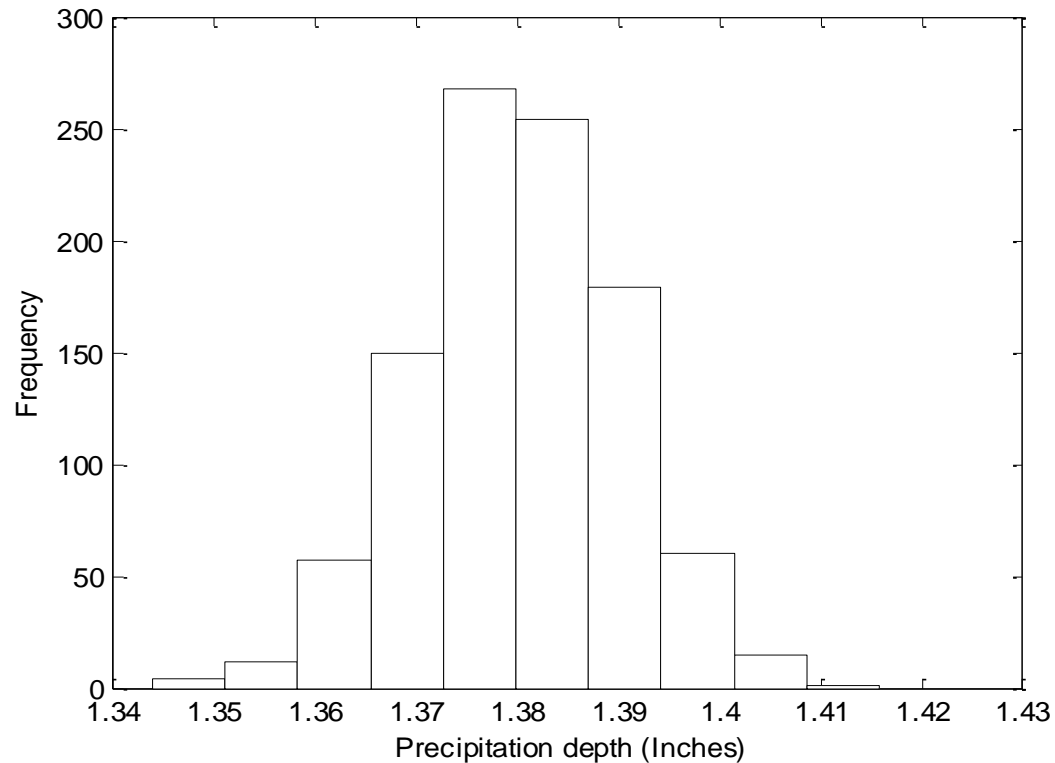
Performance Measure

	Exponent				
	1	2	4	6	8
Data Set	Absolute Error(inches)				
1	250.146	241.553	233.433	231.045	232.146
2	261.255	261.057	265.238	270.604	275.755
3	251.415	244.354	238.358	235.835	234.716
4	251.156	245.858	242.796	242.277	242.838
5	255.208	250.328	247.793	251.667	260.251
6	251.100	245.269	240.033	237.877	238.267
7	253.892	249.682	245.177	243.225	244.184
8	248.580	241.726	238.409	241.451	247.403
9	249.774	244.084	238.824	237.010	237.307
10	256.096	251.258	247.838	246.758	246.332
11	246.271	238.746	237.591	242.057	247.236
12	250.237	242.831	235.719	234.333	236.341
13	250.216	242.946	236.533	233.767	233.510
14	249.949	240.993	232.530	228.759	226.822
15	252.459	246.362	239.725	235.853	234.144
16	249.753	241.956	235.543	233.825	234.375
17	255.395	250.017	247.616	249.459	252.966
Average	251.935	245.825	241.362	240.929	242.623

Under and Over Estimation

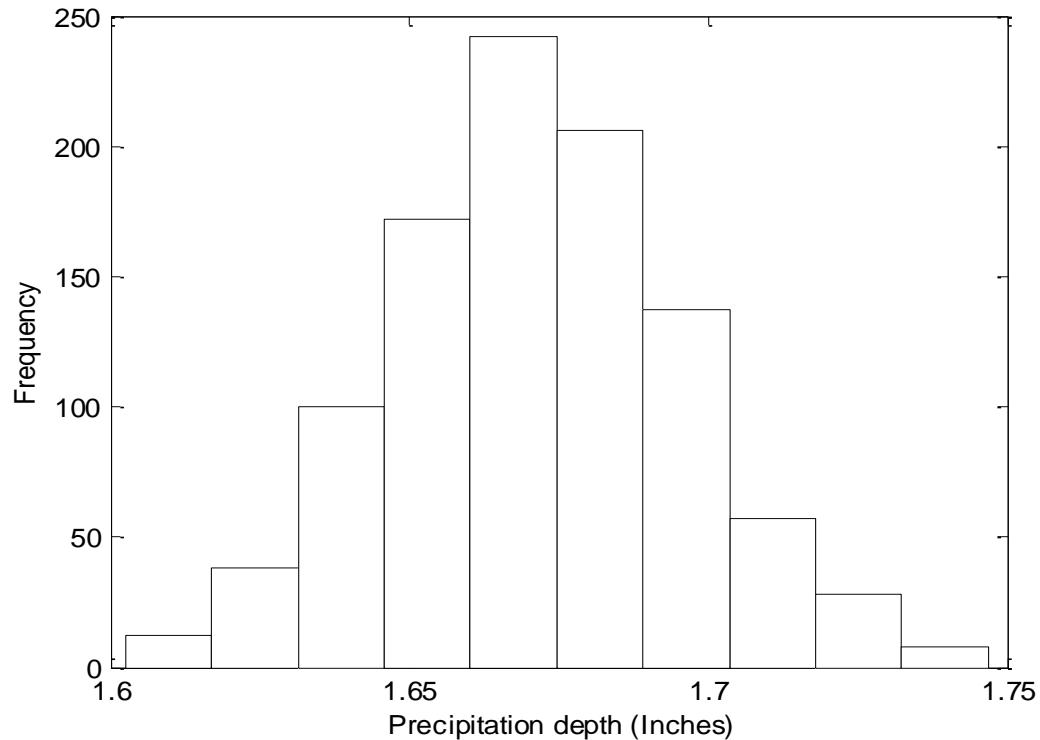


Distribution of Estimates



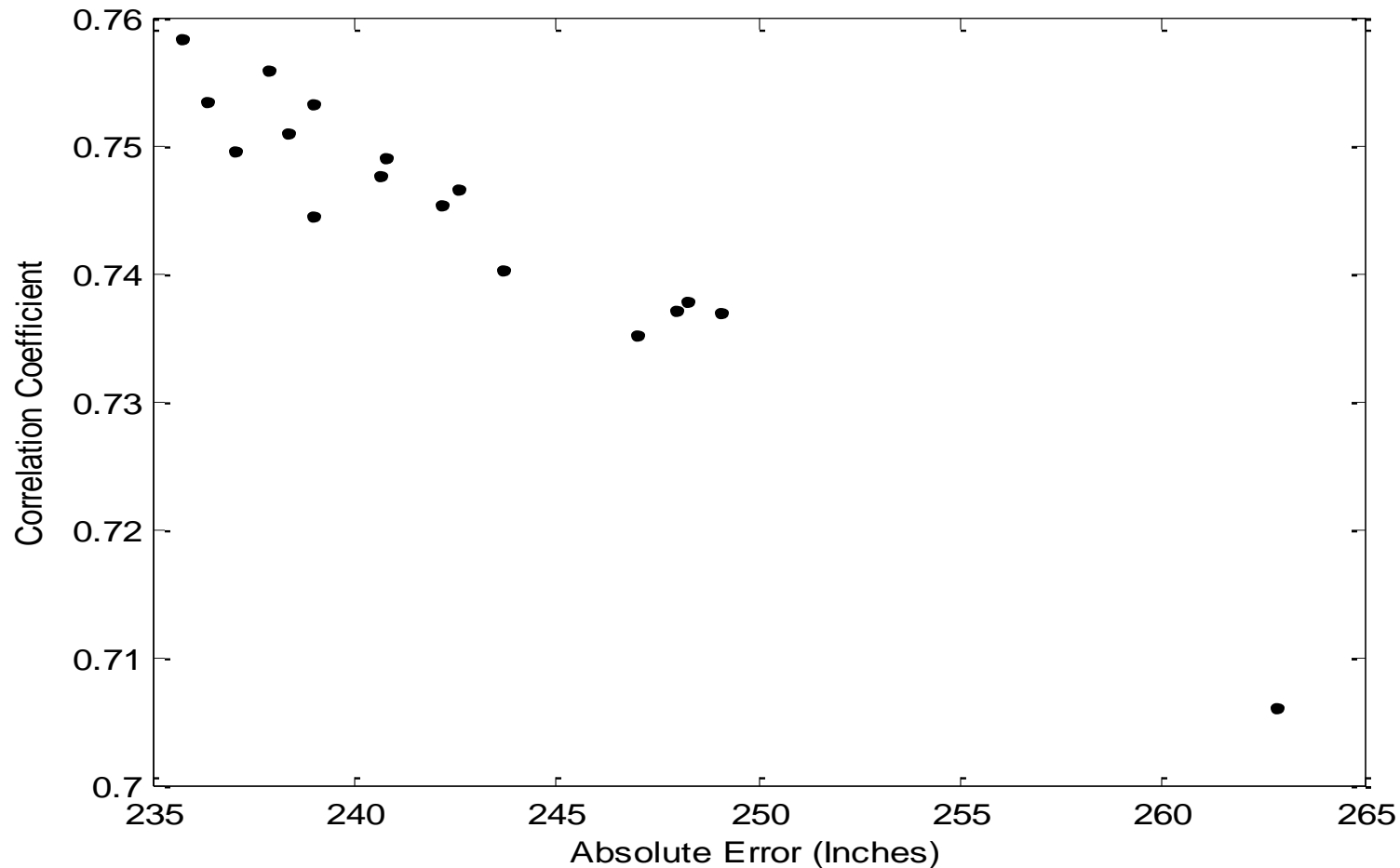
Histogram of estimated missing precipitation values for January 6, 1995 (observed value: 1.38 inches) using multiple Imputation

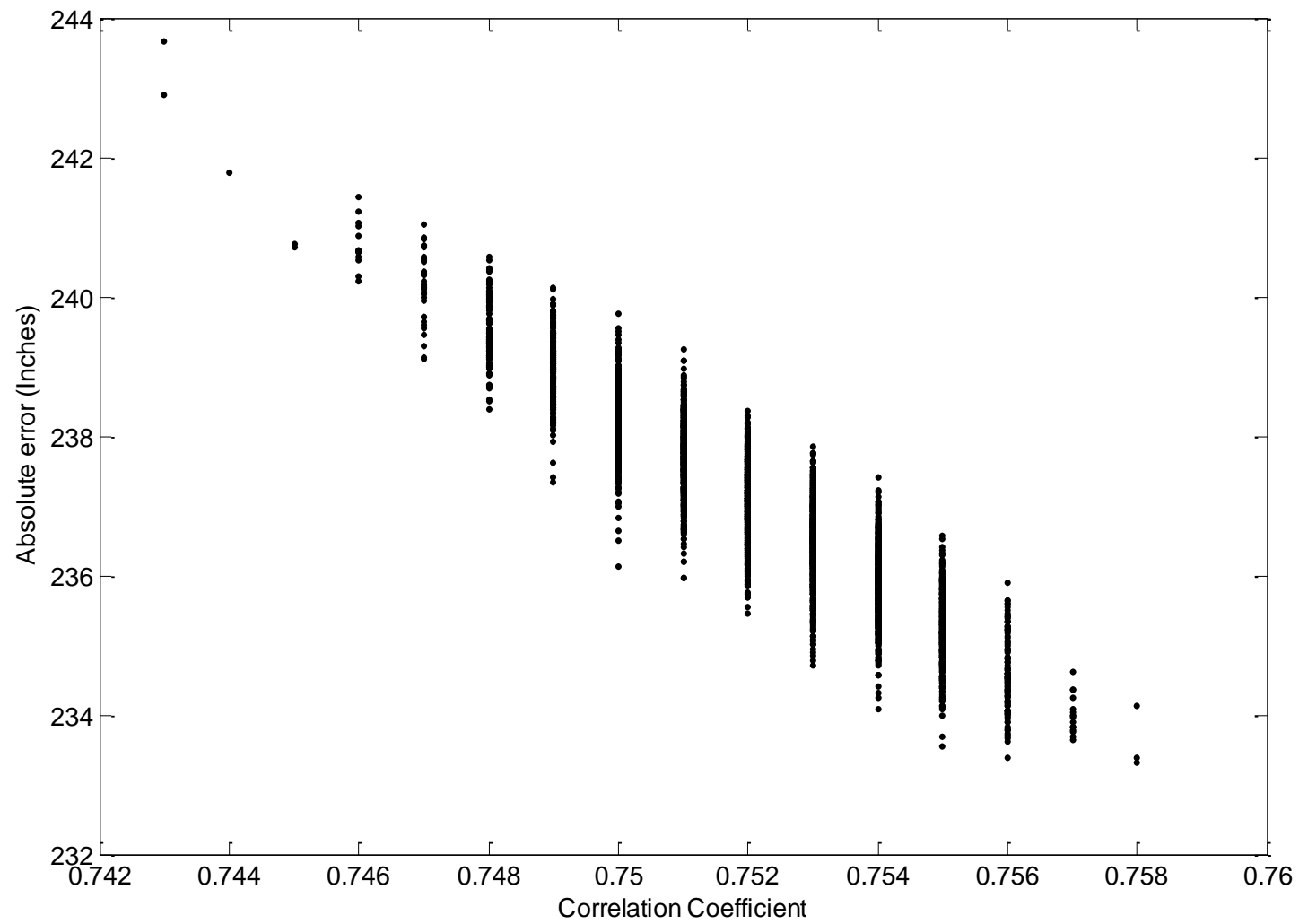
Distribution of Estimates



Histogram of estimated missing precipitation values for
May 18, 1995 (observed value: 1.68 inches) using bootstrap re-sampling

Joint Variability of Performance Measures





Observations

- Imputed data (based on single imputation) may introduce heavy biases into evaluation of trends in historical data (Teegavarapu, 2013)
- Complete data sets can be evaluated by Imputation, Pooling and Assessment (IAP) concepts.
- Total variance from the completed data sets can be obtained by within imputation variance and between imputation variance.

Observations

- **Multiple imputation methods are used for estimating missing precipitation data at station using resampling schemes embedded in optimal weighting methods.**
- **Random sampling of neighborhood points without replacement to form a set of control points and bootstrap resampling schemes are used to sample the spatio-temporal precipitation data enable the creation of multiple filled data sets.**
- **The proposed methodology with a specific re-sampling scheme and interpolation provides a mechanism to obtain error estimates for each estimated value of precipitation.**

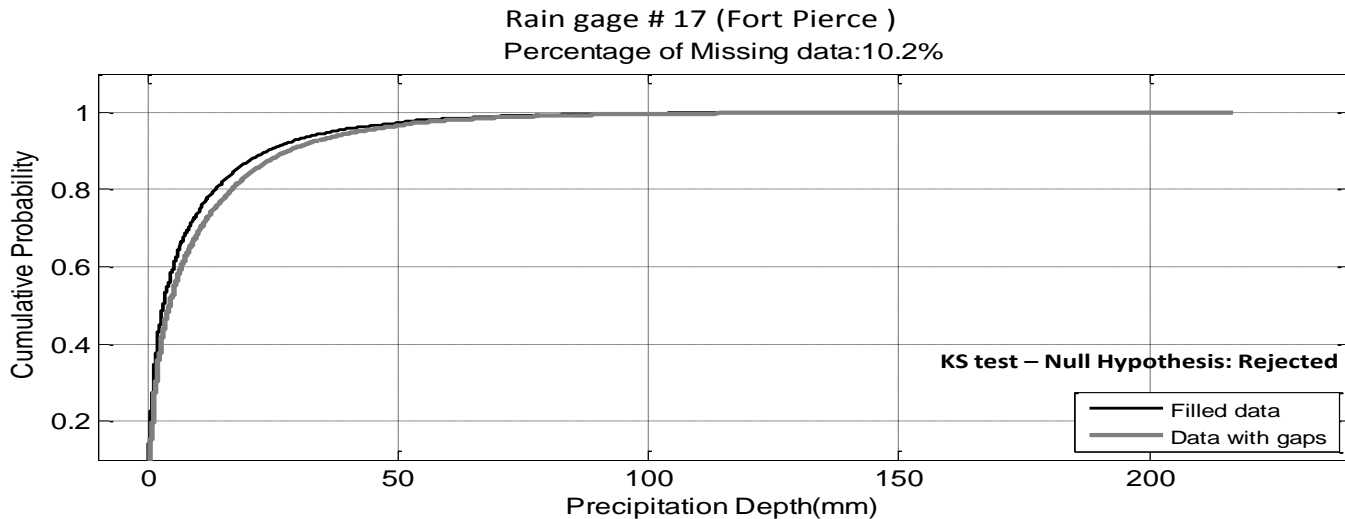
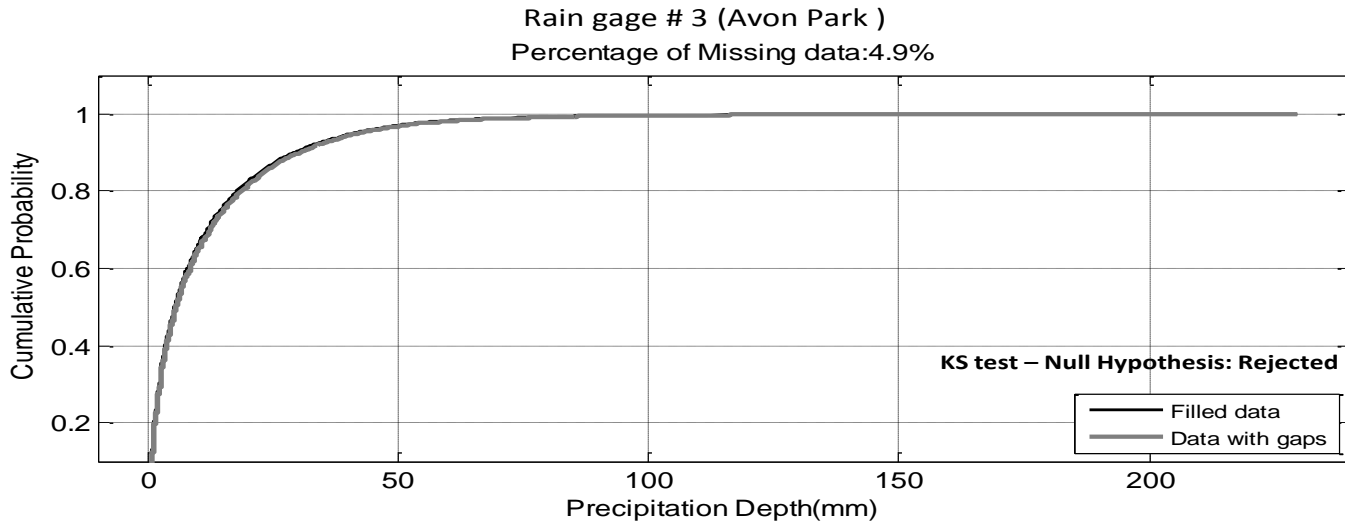
-
- **Mathematical programming models using nonlinear and mixed integer nonlinear programming (MINLP) formulations along with binary variables are used for solving the optimal spatial interpolation. The binary variables are used to select the optimal number of stations (best estimators) for revised estimates of missing precipitation data from already interpolated data.**
 - **Post-interpolation bias-corrections will preserve the site-specific summary statistics with minor changes in the magnitudes of error and performance measures.**
 - **Multi-model imputation and combined and spatial and temporal resampling methods are not evaluated in this study. However, multi-model imputation and spatial neighborhood (non optimal) variability will result in higher variability in estimates compared to temporal resampling approach.**

Issues with Filled Datasets

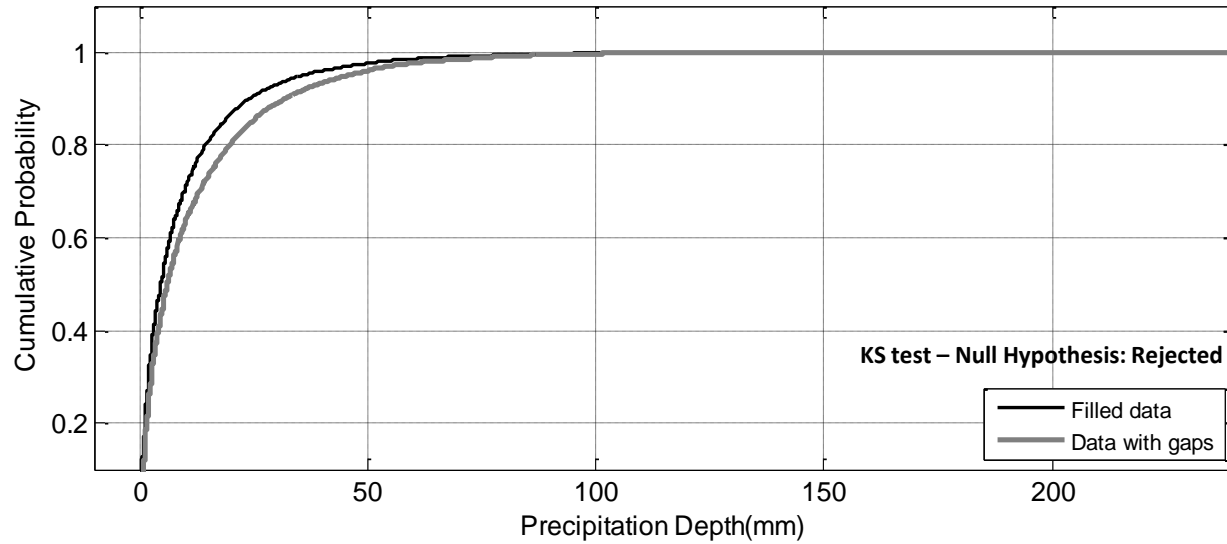
- Data infilling **may not** introduce statistically significant bias in precipitation values at large temporal resolution.
- Infilling however may lead to underestimation of both magnitude and frequency of heavy and very heavy precipitation events.
- infilling may also affect the spatial characteristics of extreme precipitation in the region.
- Bias introduced by the data infilling increases as gaps (i.e. amounts of missing data) in precipitation data increase.
- Therefore, care should be taken while analyzing extreme precipitation events from precipitation data where gaps have been infilled.

Variations in CDFs

Length of Filled Data

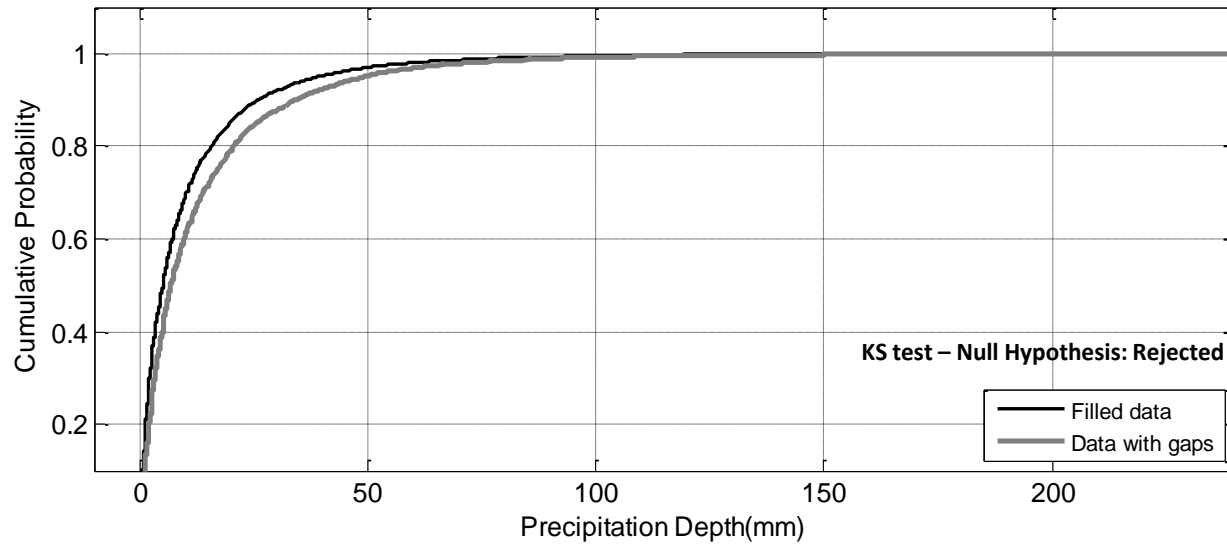


Rain gage # 31 (Lake Alfred)
Percentage of Missing data:29.7%

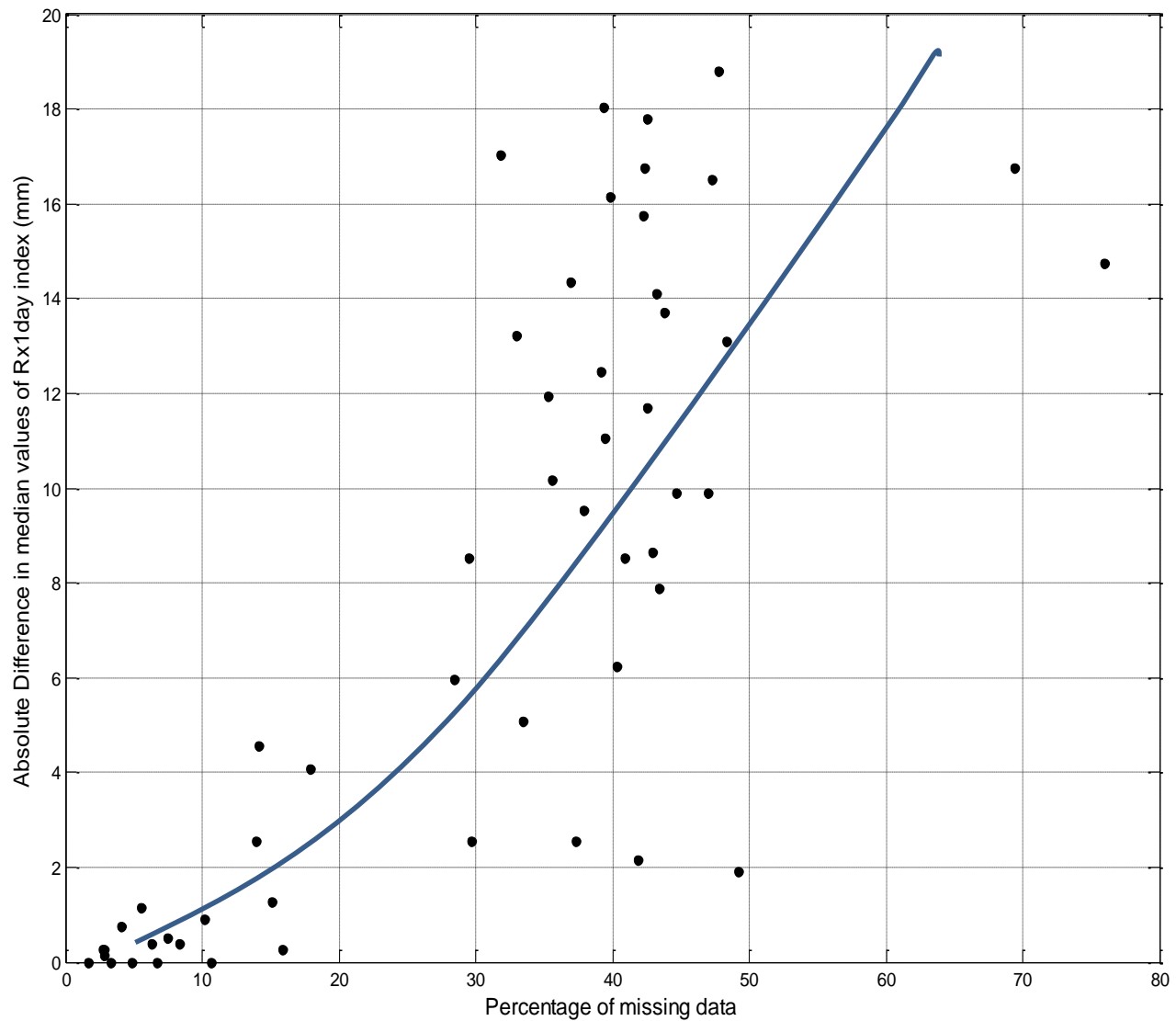


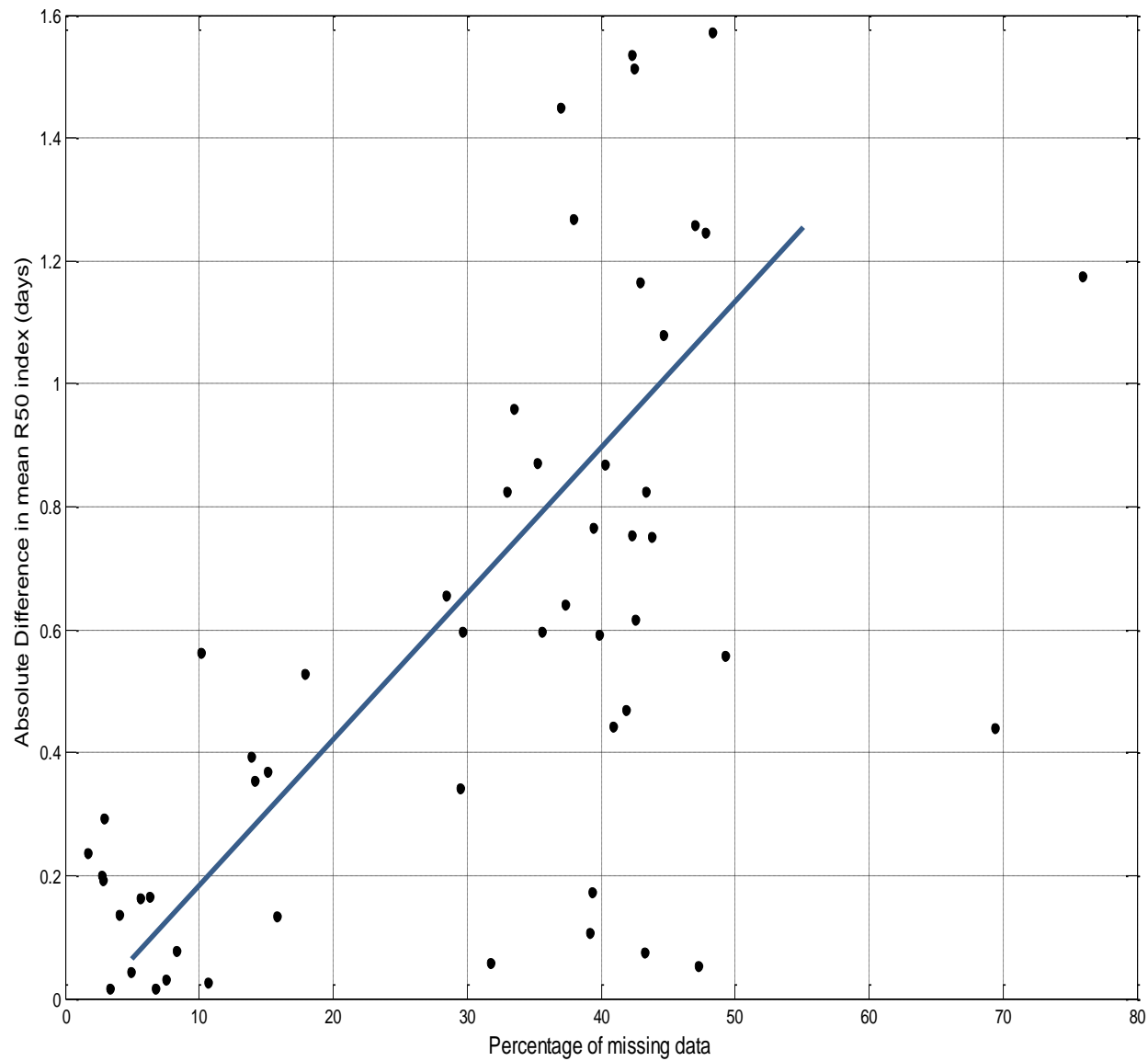
(c)

Rain gage # 23 (Hypoluxo)
Percentage of Missing data:49.2%



(d)





Minimize

$$W_{RMSE} \sqrt{\frac{\sum_{n=1}^{no} (\theta_{m,n}' - \theta_{m,n})^2}{no}} + W_{\mu} |\mu_m - \dot{\mu}_m| + W_{\sigma} |\sigma_m - \dot{\sigma}_m| + W_{\sigma*} \sum_{i=1}^{ns-1} (|\sigma_i - \dot{\sigma}_m| -$$

Variable	Explanation
$\theta_{m,n}$	Observed values at station m
$\theta_{m,n}'$	Estimated value of precipitation at base station, m
μ_m	Mean of series based on observed values
$\dot{\mu}_m$	Mean of combine data series (observed and estimated values)
σ_m	Standard deviation of series based on observed values
$\dot{\sigma}_m$	Standard deviation of combine data series based on observed and estimated values
σ_i	Standard deviation of precipitation data series at station i
μ_i	Mean of precipitation data series at station i
$\dot{\rho}_{i,m}$	Correlation coefficient based on data series at station i and station m (observed and estimated values)
$\rho_{i,m}$	Correlation coefficient based on data series at station i and station m based on observed values
$W_{RMSE}, W_{\mu}, W_{\sigma}, W_{\sigma*}, W_{\mu*}, W_{\rho*}$	User specified weights for each objective

Balance trade offs between minimizing
estimation errors and preserving
site and regional statistics

Original formulation

(Focus on Estimation Error)

$$\textbf{maximize: } f(x) = C^T x$$

subject to:

$$Ax \leq b$$

$$Dx \leq b'$$

$$x \geq 0$$

$$c, x \in \Re^n, b \in \Re^m, A \in \Re^{m \times n}$$

Intermediate formulation

(Tolerate Loss in
Estimation Error Performance)

$$\textbf{maximize: } f(x) = C^T x$$

subject to:

$$Ax \leq b + t_0$$

$$Dx \leq b'$$

$$x \geq 0$$

Final formulation

maximize: L

subject to:

$$L(f_1 - f_0) + C^T x \leq f_1$$

$$Lt_0 + Ax \leq b + t_0$$

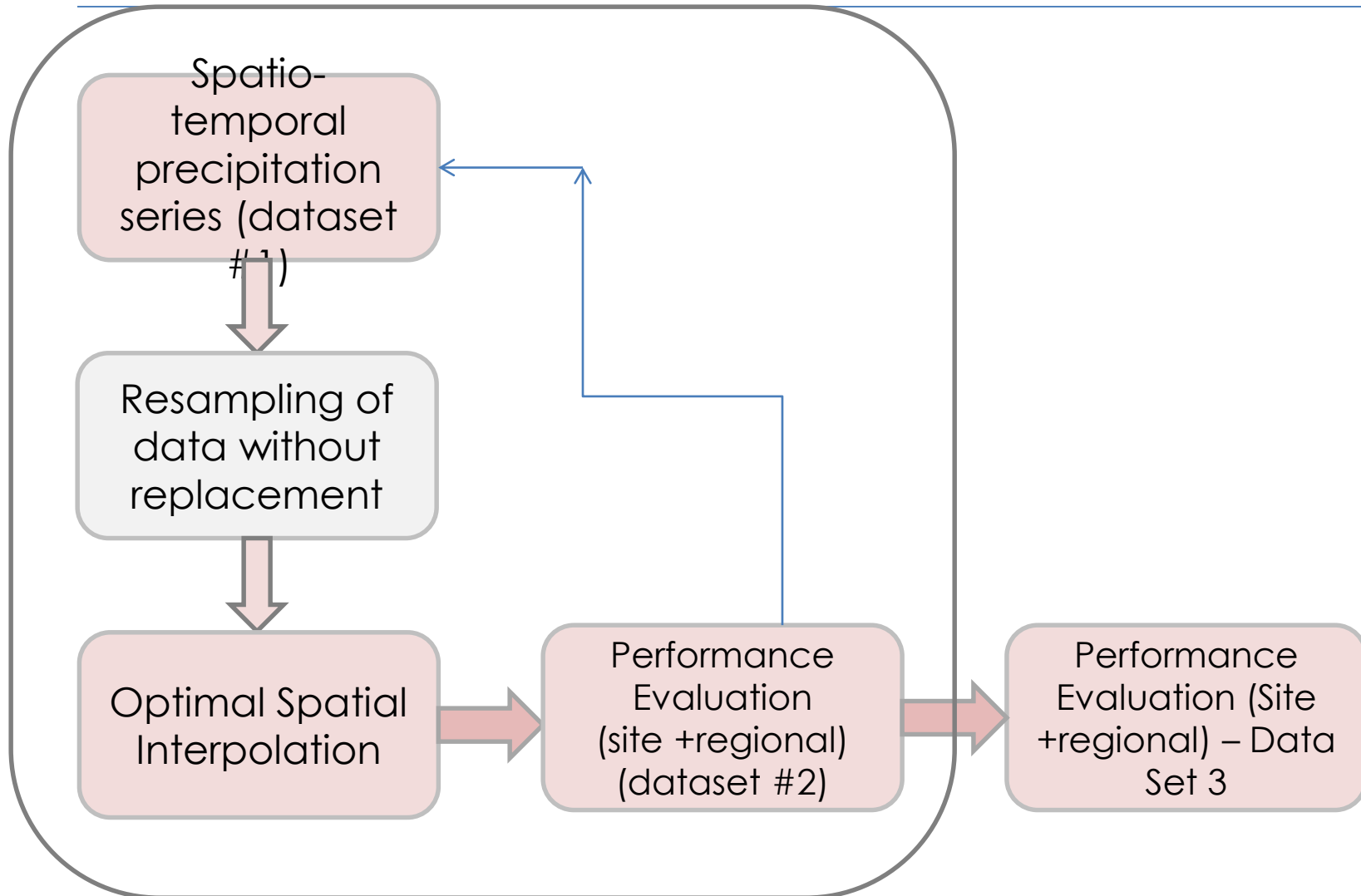
$$Dx \leq b'$$

$$L, x \geq 0$$

$$L \leq 1$$

$$L = \begin{cases} 1, & \text{if } f_0 < C^T x \\ \frac{C^T x - f_1}{f_0 - f_1}, & \text{if } f_1 < C^T x < f_0 \\ 0, & \text{if } C^T x \leq f_1 \end{cases}$$

Resampling Optimization Method



Single/Multiple Best Estimators for Corrections

Minimize $\sqrt{\frac{\sum_{i=1}^{no} (\theta_{\alpha,i}^{ce} - \theta_{\alpha,i}^o)^2}{no}} \quad \forall \alpha$

$$\tau_{\alpha,i} = \sum_{j=1}^{ns-1} \theta_{j,i}^o \lambda_j \quad \forall \alpha, \forall i$$

$$\tau_{\alpha,i} \leq N \beta \quad \forall \alpha, \forall i$$

$$\theta_{\alpha,i}^{ce} = \theta_{\alpha,i}^e \beta \quad \forall \alpha, \forall i$$

$$\sum_{j=1}^{ns-1} \lambda_j = \phi$$

$$\sum_{j=1}^{ns-1} \lambda_j \leq \phi$$

λ_j and β are binary variables

N is a large number

no is the number of time intervals,

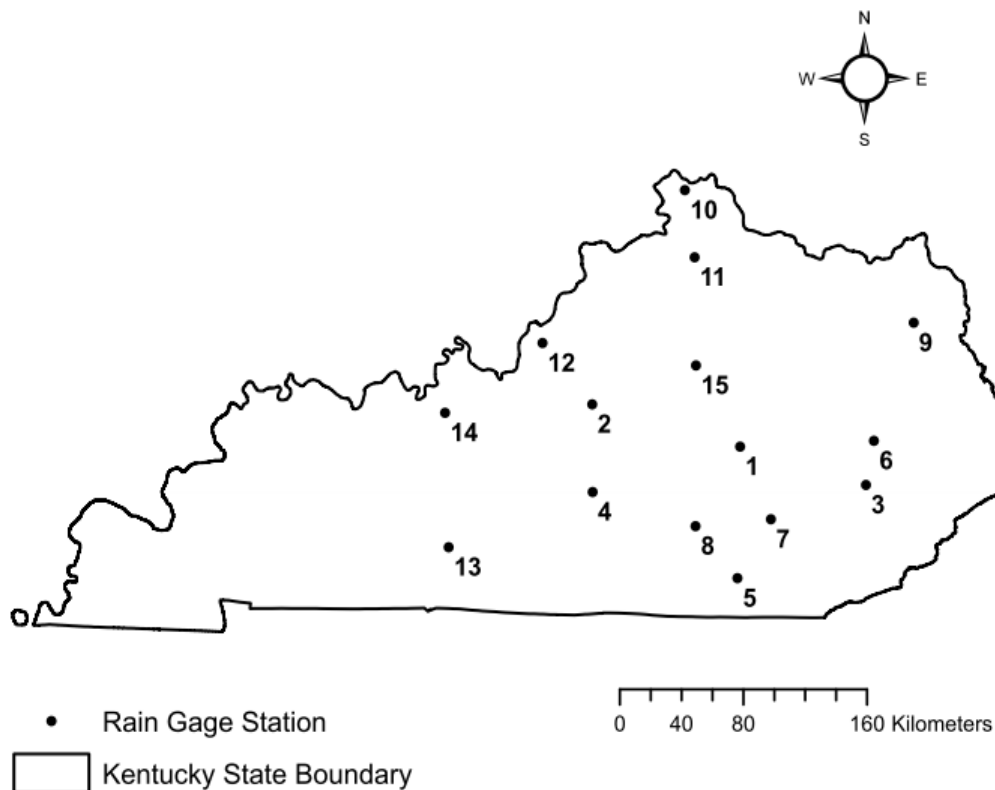
i is the index for time interval,

α is base rain gauge

$\theta_{\alpha,i}^o$ is the observed value

ϕ is an integer used to define the number of estimators

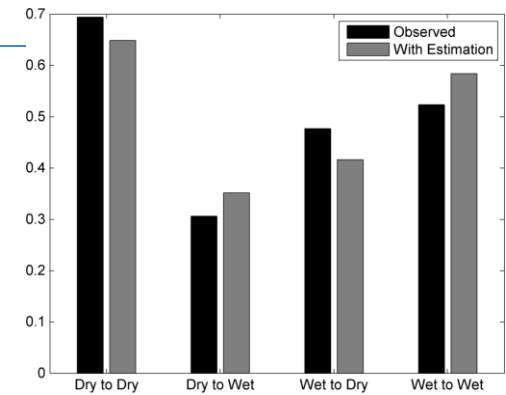
Case Study Application



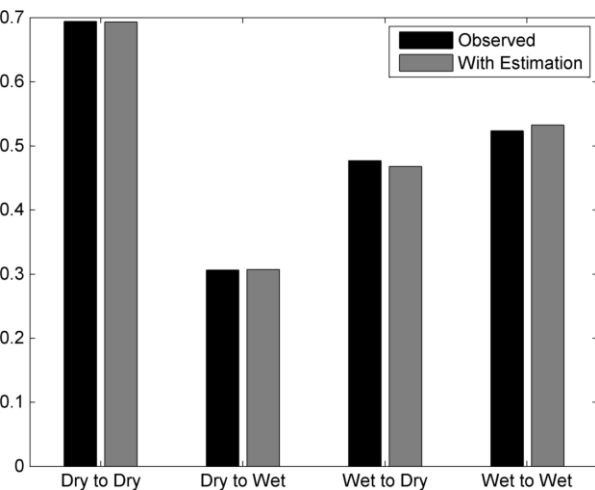
The state has a temperate climatic zone in which the majority of storm events are due to frontal precipitation.

The Köppen-Geiger climate classification scheme (Kottek et al., 2006) for the state of Kentucky is referred to as “*Cfa*”. The climate is classified as warm temperate (*C*), fully humid (*f*) and hot summer (*a*).

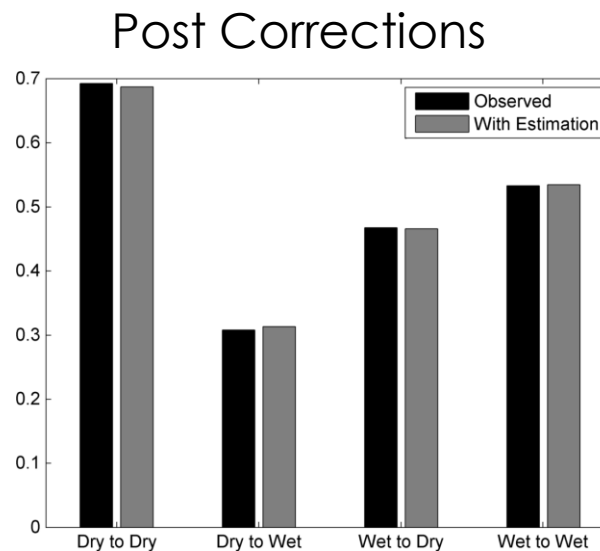
The post correction of interpolation methods preserved at site transition probabilities.



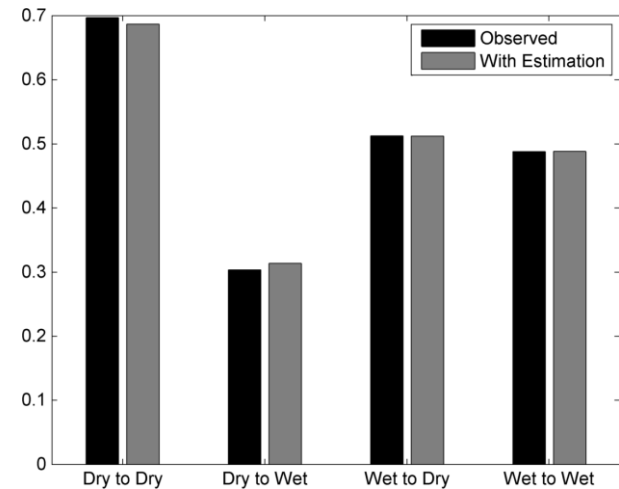
Optimal
interpolation
no post-correction



Station 1

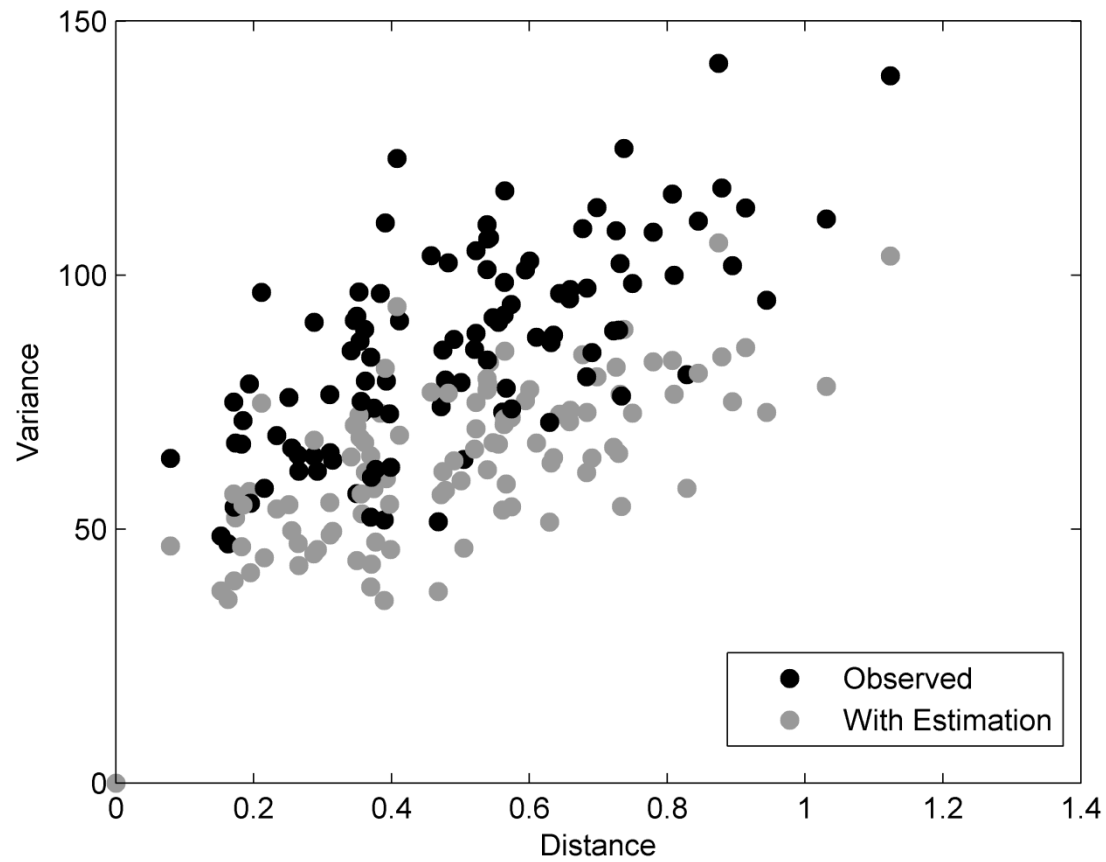


Station 7

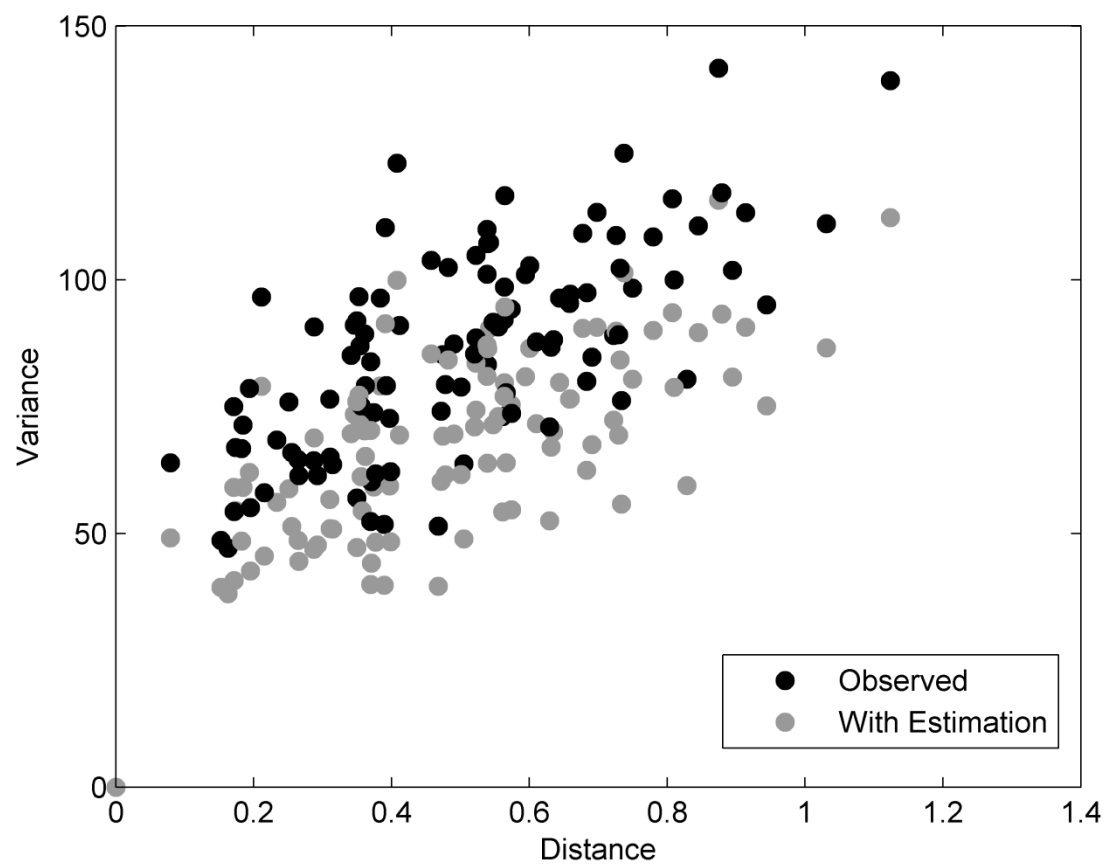


Station 12

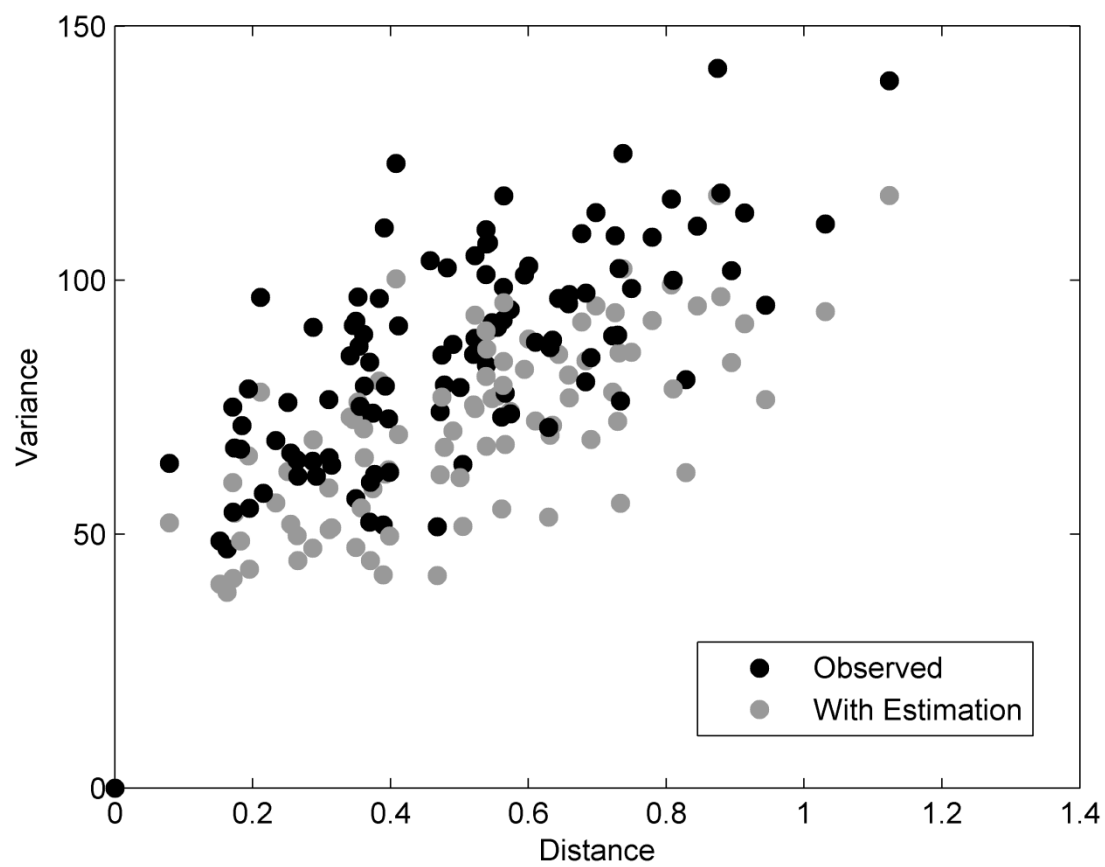
Spatial Variance



Gage Mean Method



SOFW, RMSE



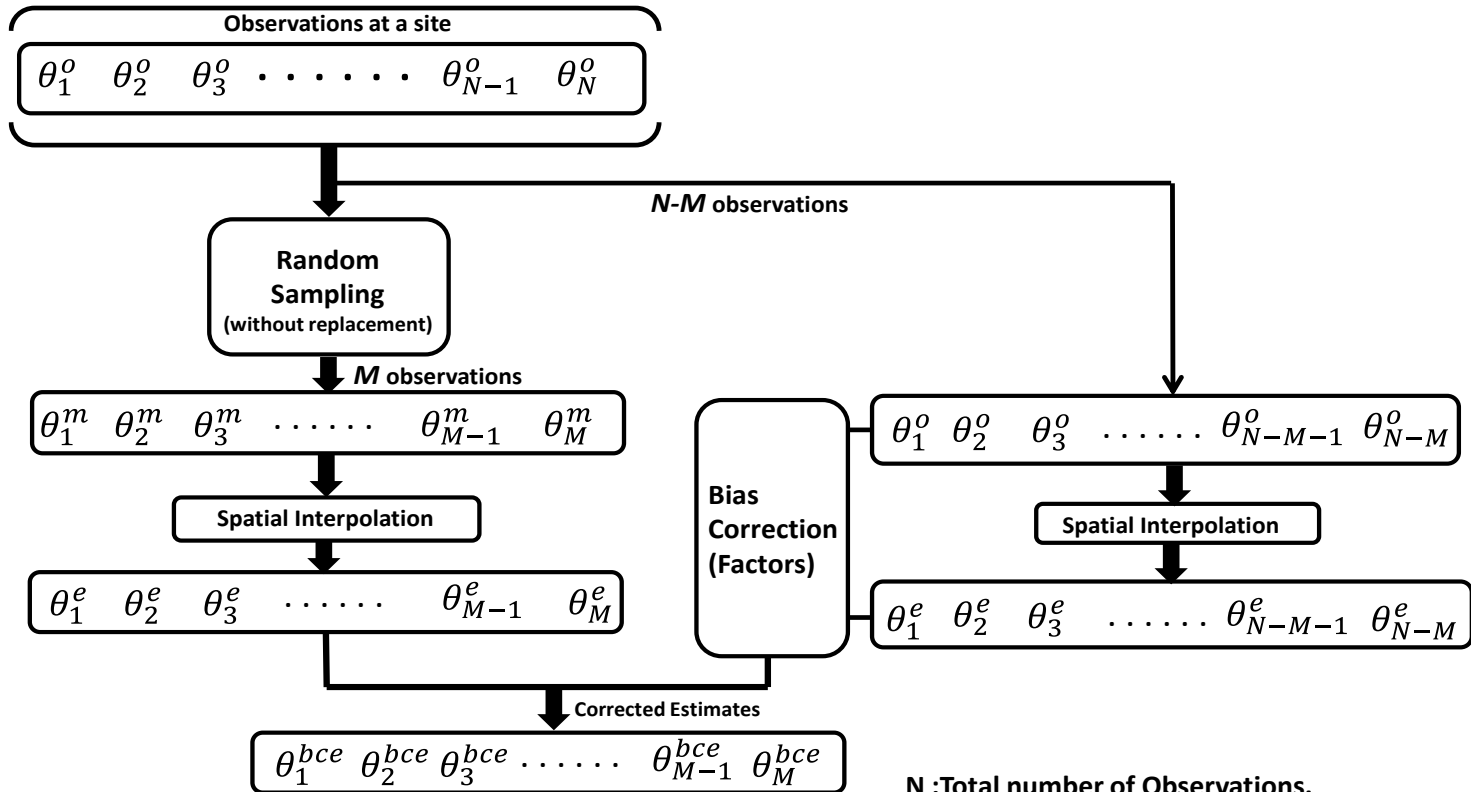
Fuzzy Logic

-
- All stochastic and deterministic spatial interpolation methods have major limitation of under estimating higher end extremes and over estimating lower end extremes.
 - New multi-objective optimal spatial interpolation methods are proposed and evaluated.
 - Optimal single and multiple best estimators for dry and wet event corrections used after the multi-objective methods improved the estimation:
 - Preservation of site-specific statistics
 - Preservation of regional-to-site relationships
 - Decreased the inflation of site-to-regional correlations
 - Brought autocorrelation structure and transition probabilities close to those from the observed series.
 - Improved spatial variability (brought the variability close to existing)
 - Balancing trade-offs between Estimation Error Performance, Site and Regional Statistics with new interpolation techniques

Statistical Corrections of Spatially Interpolated Estimates

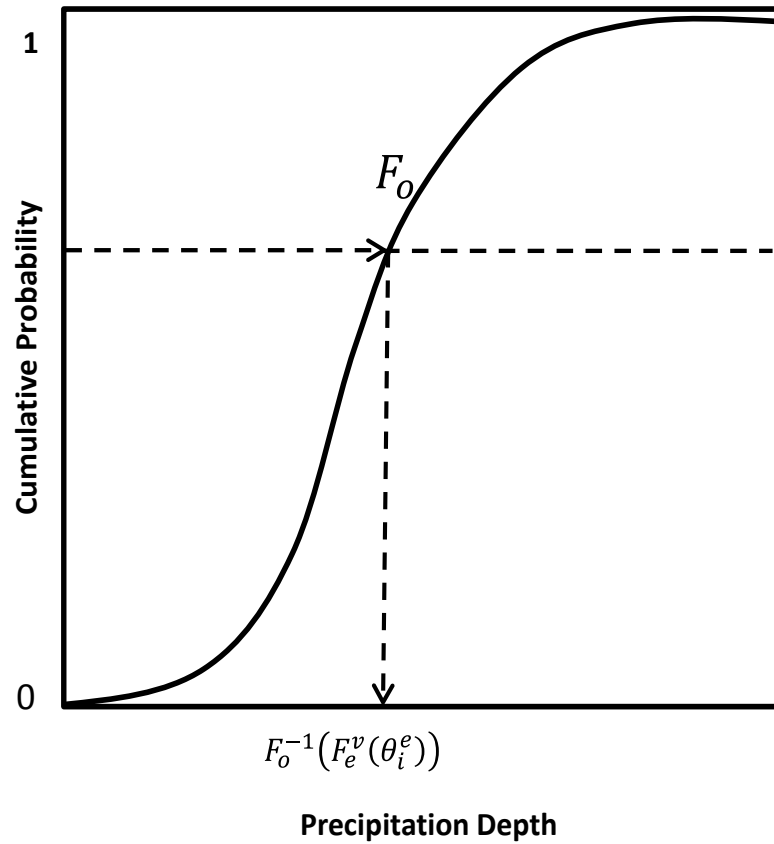
- Spatial interpolation methods used for estimation of missing precipitation data generally under and overestimate the high and low extremes respectively.
- This is a major limitation that plagues all spatial interpolation methods as observations from different sites are used in local or global variants of these methods for estimation of missing data.
- A bias-correction methods similar to those used in climate change studies for correcting missing precipitation estimates provided by an optimal spatial interpolation method.
- The methods are applied to post-interpolation estimates using quantile-mapping, a variant of equi-distant quantile-matching and a new optimal single best estimator (SBE) scheme.

Methodology

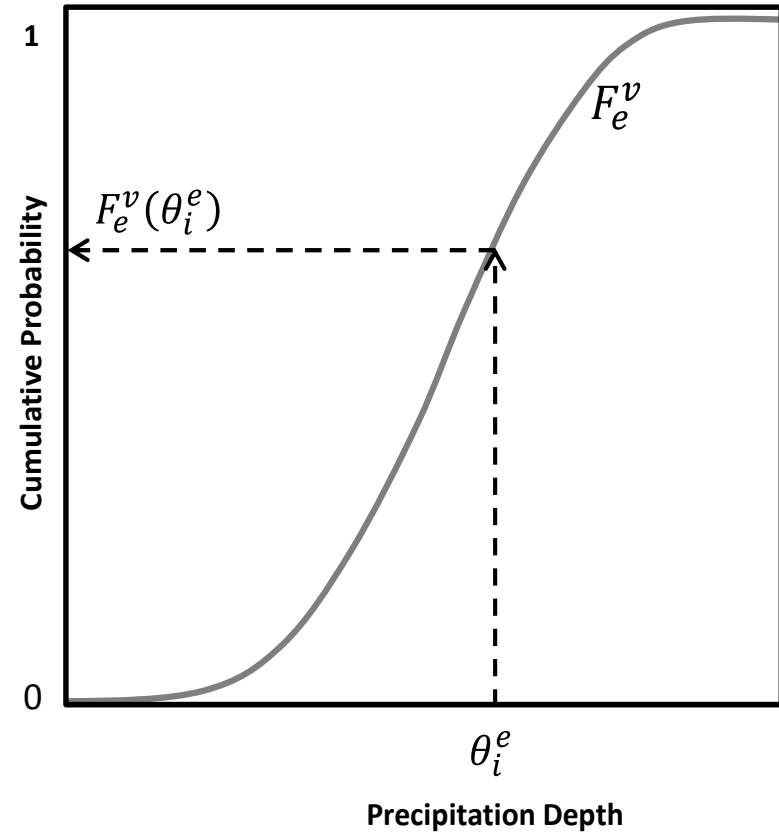


N :Total number of Observations.
M :Number of random samples.
 θ_i^m :Observation from random sample assumed to be missing.
 θ_i^e : Estimated value using interpolation.
 θ_i^{bce} : Bias corrected value using estimated value from interpolation.

Quantile Matching - Corrections

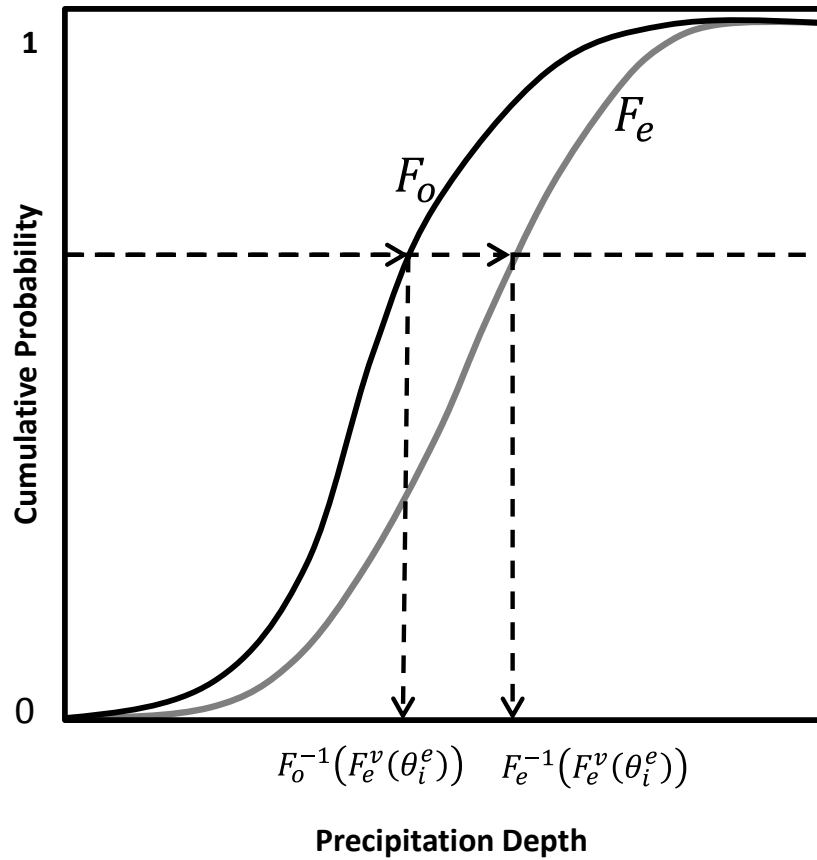


(a)

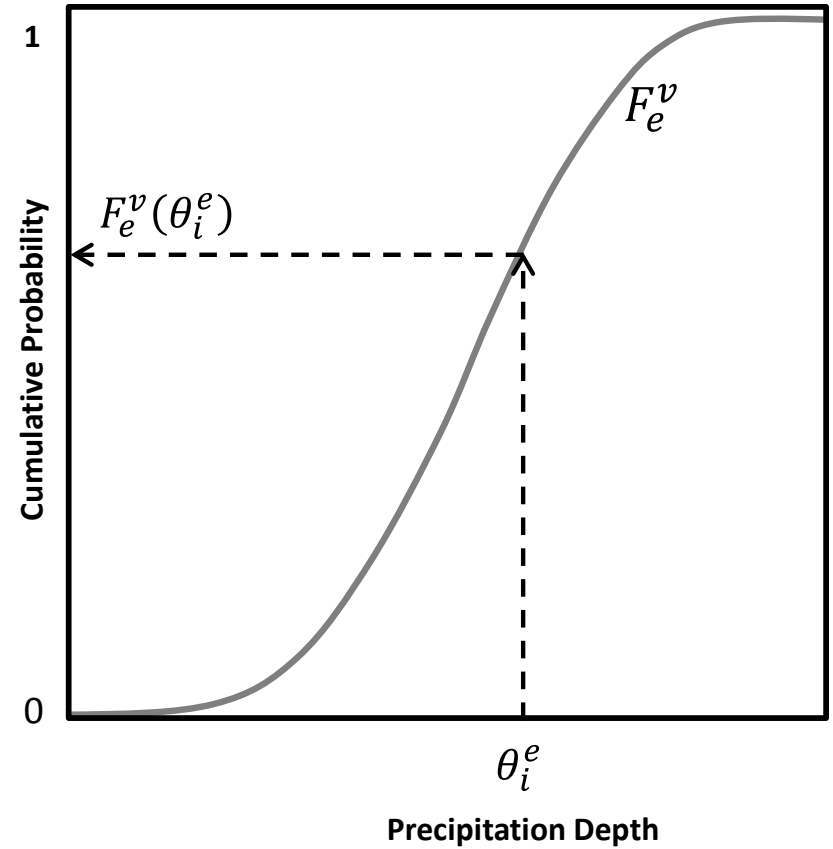


(b)

Equi-ratio Quantile Matching

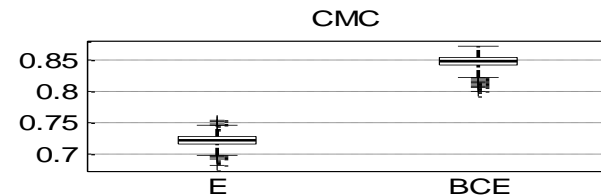
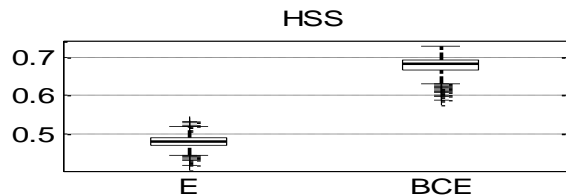
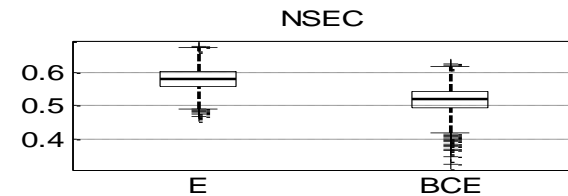
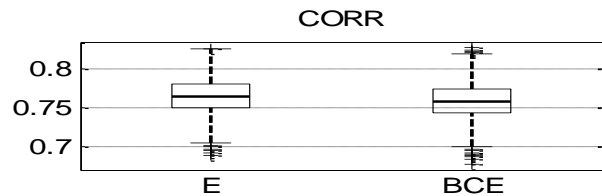
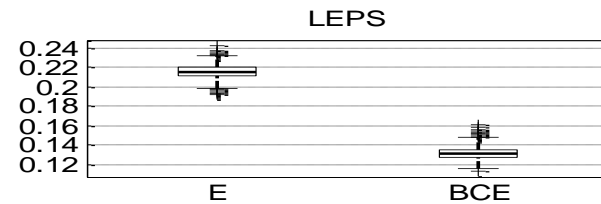
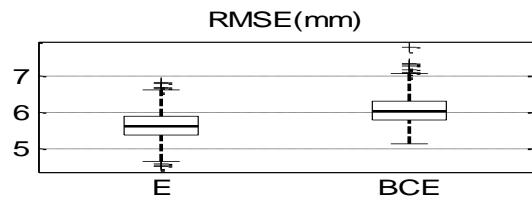
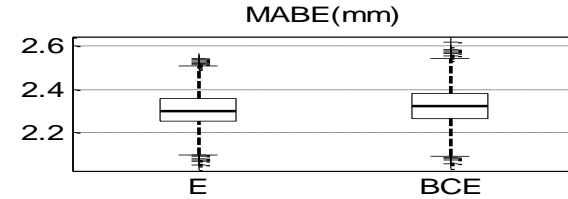
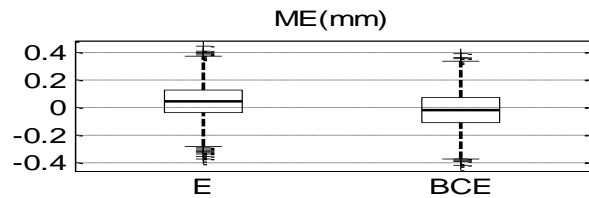


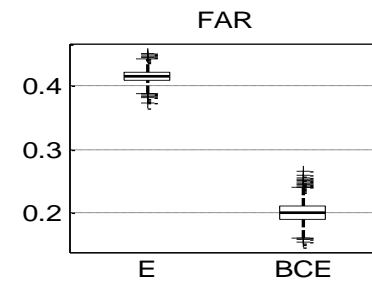
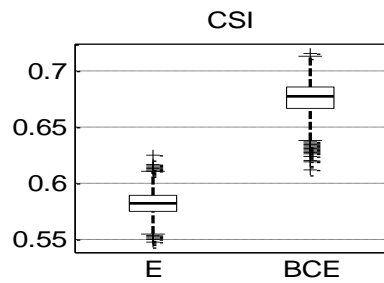
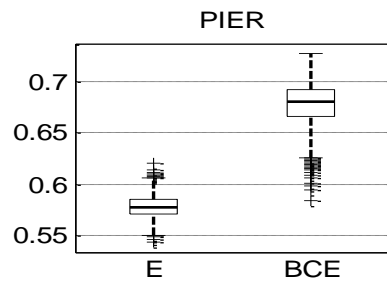
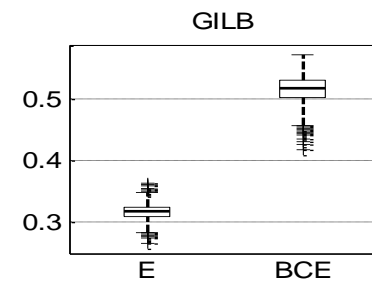
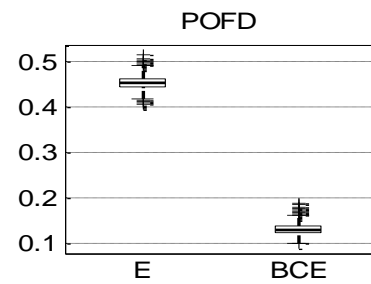
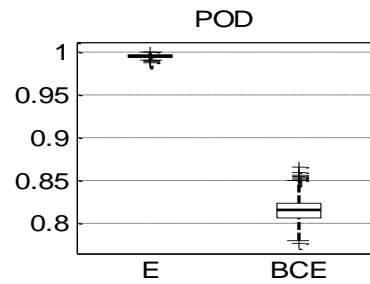
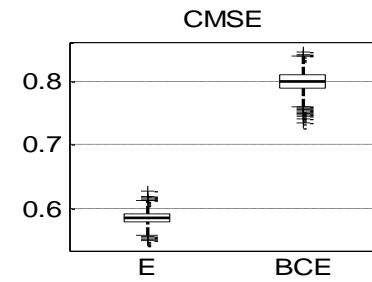
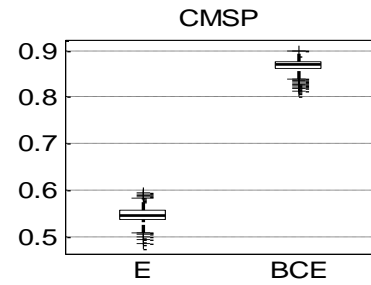
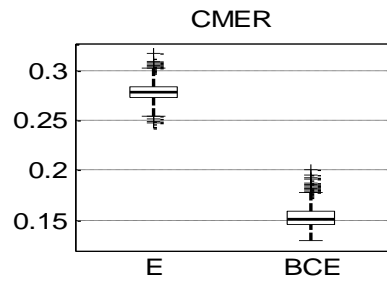
(a)



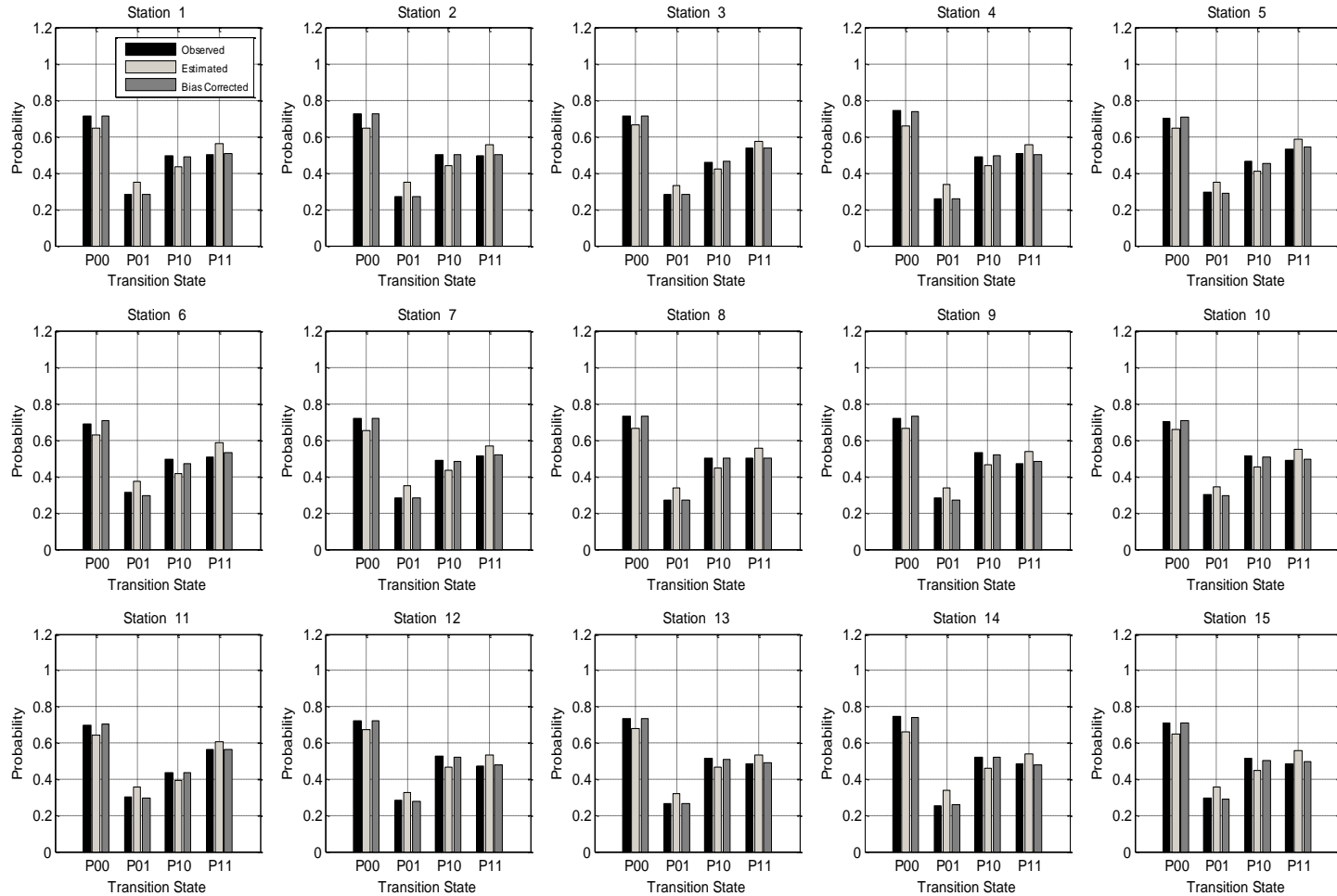
(b)

Performance and contingency measures, skill scores

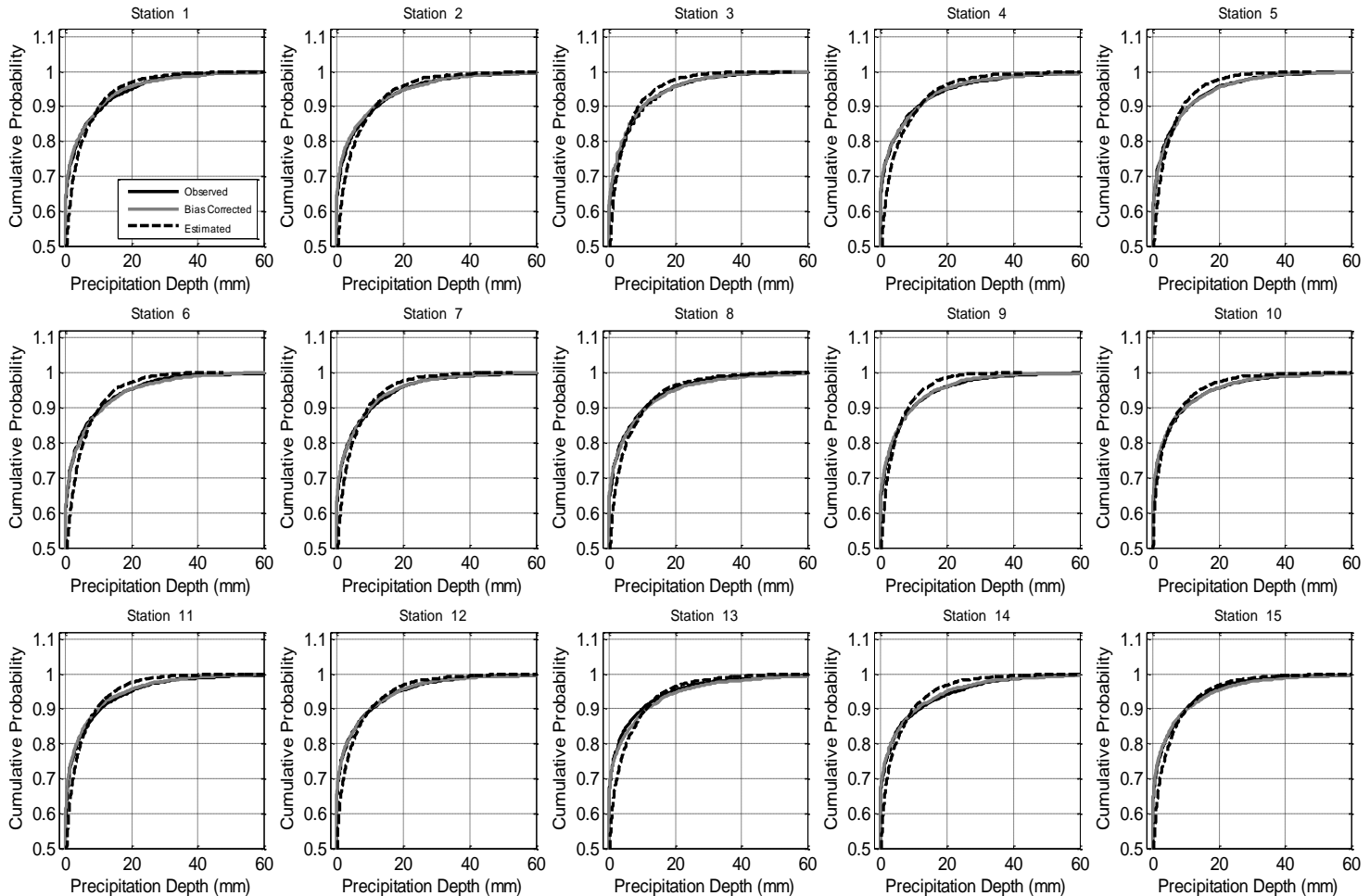




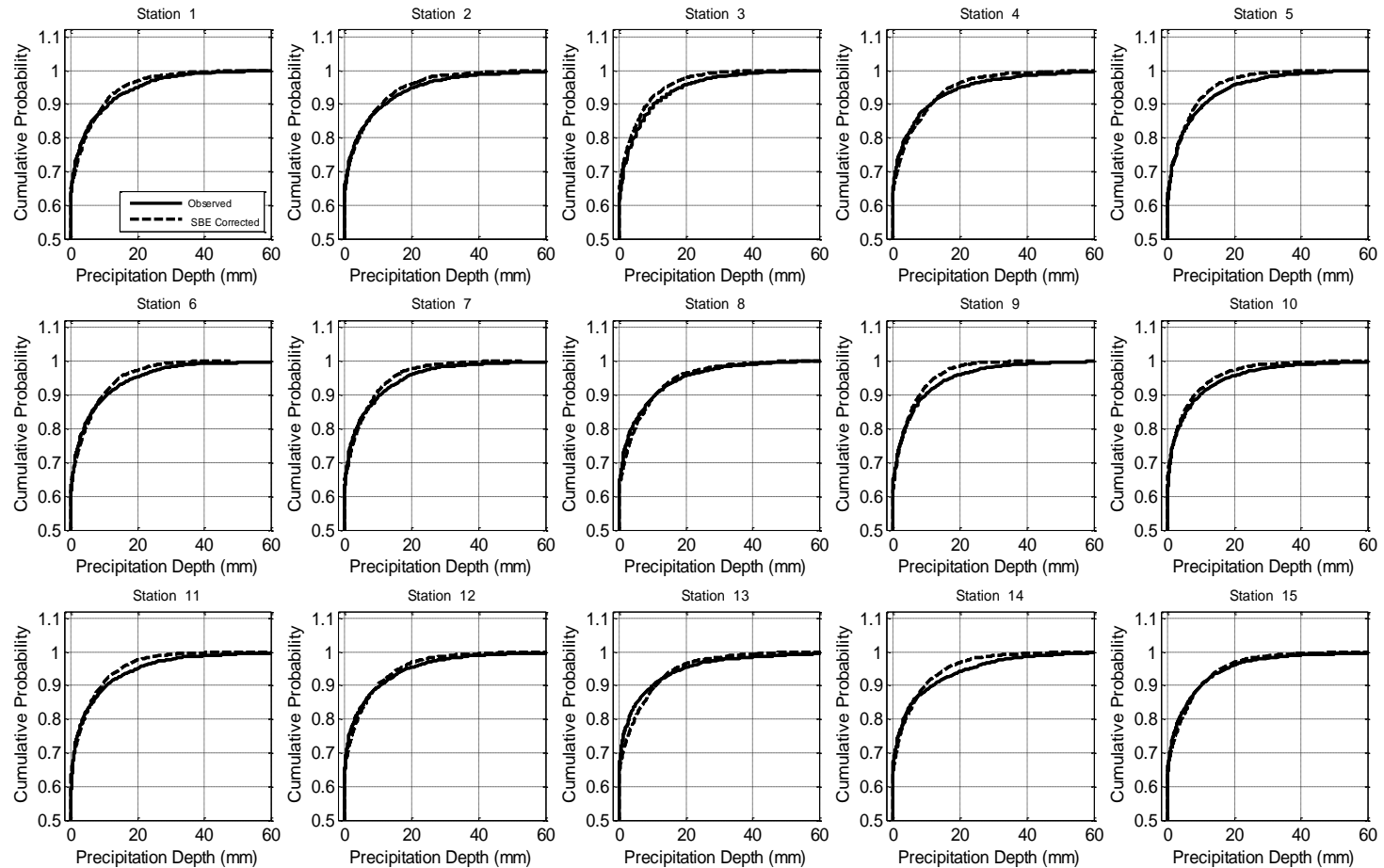
Transition Probabilities



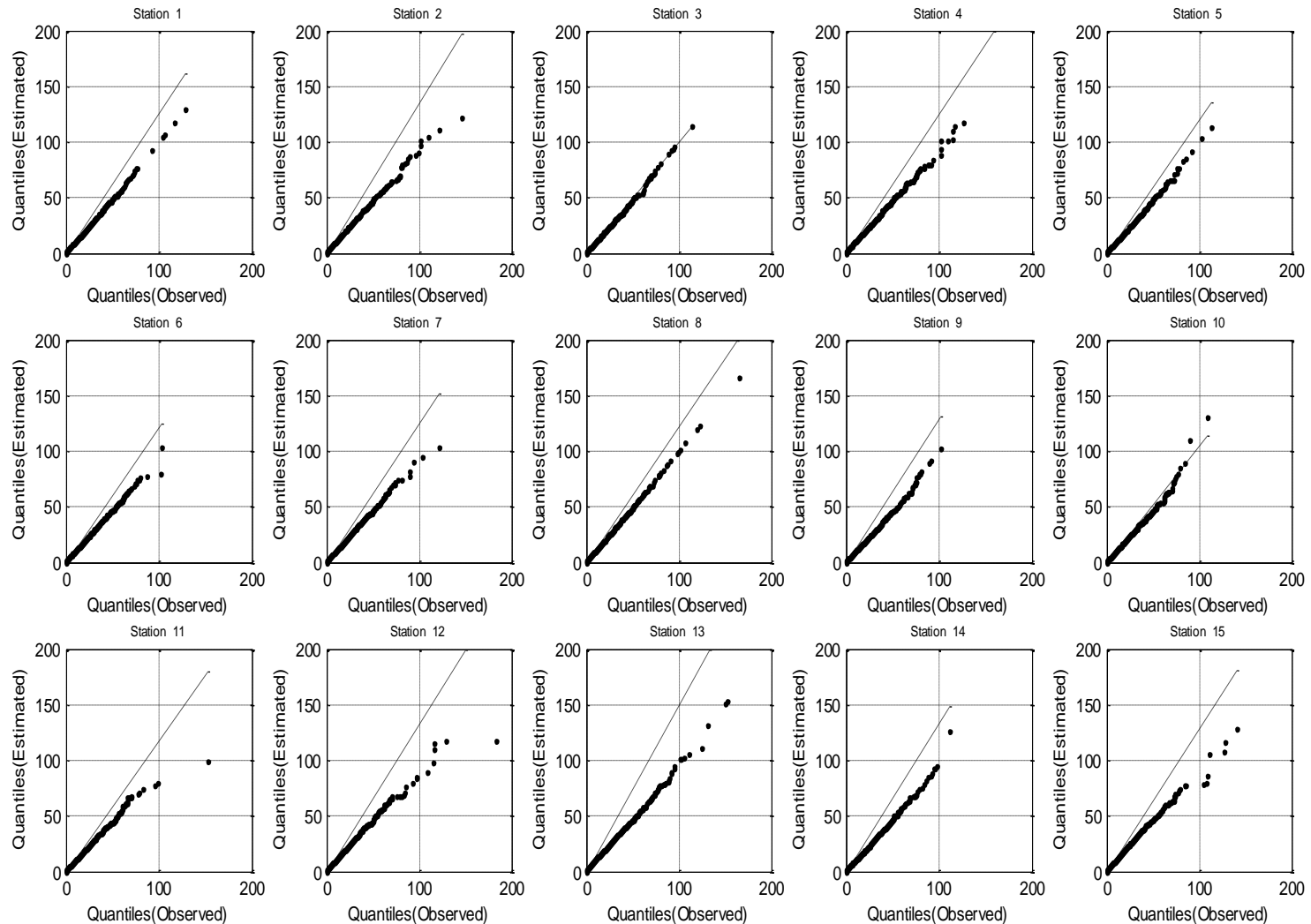
Non-exceedance Probability Curves EQRM corrected



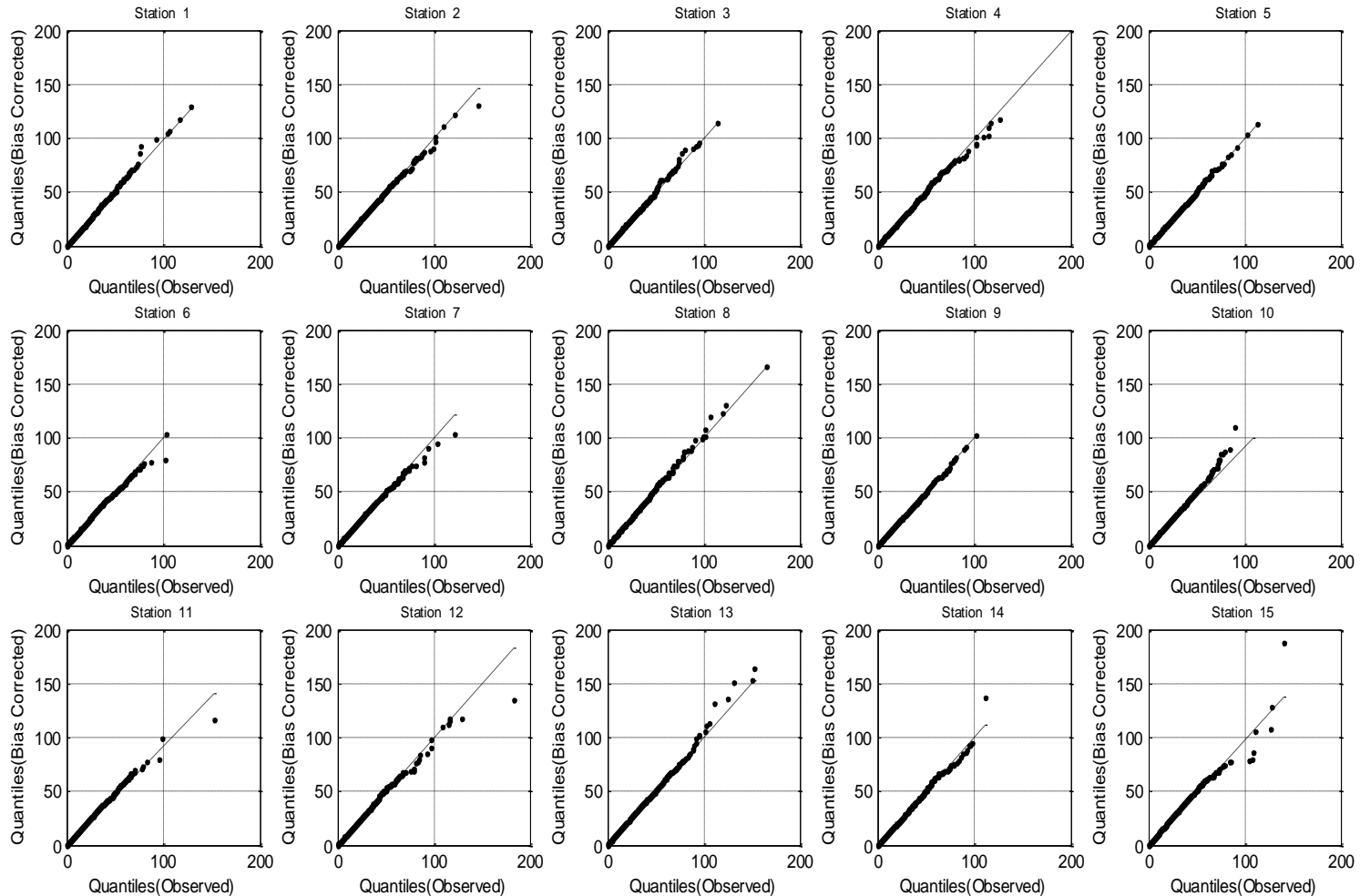
SBE-Corrected



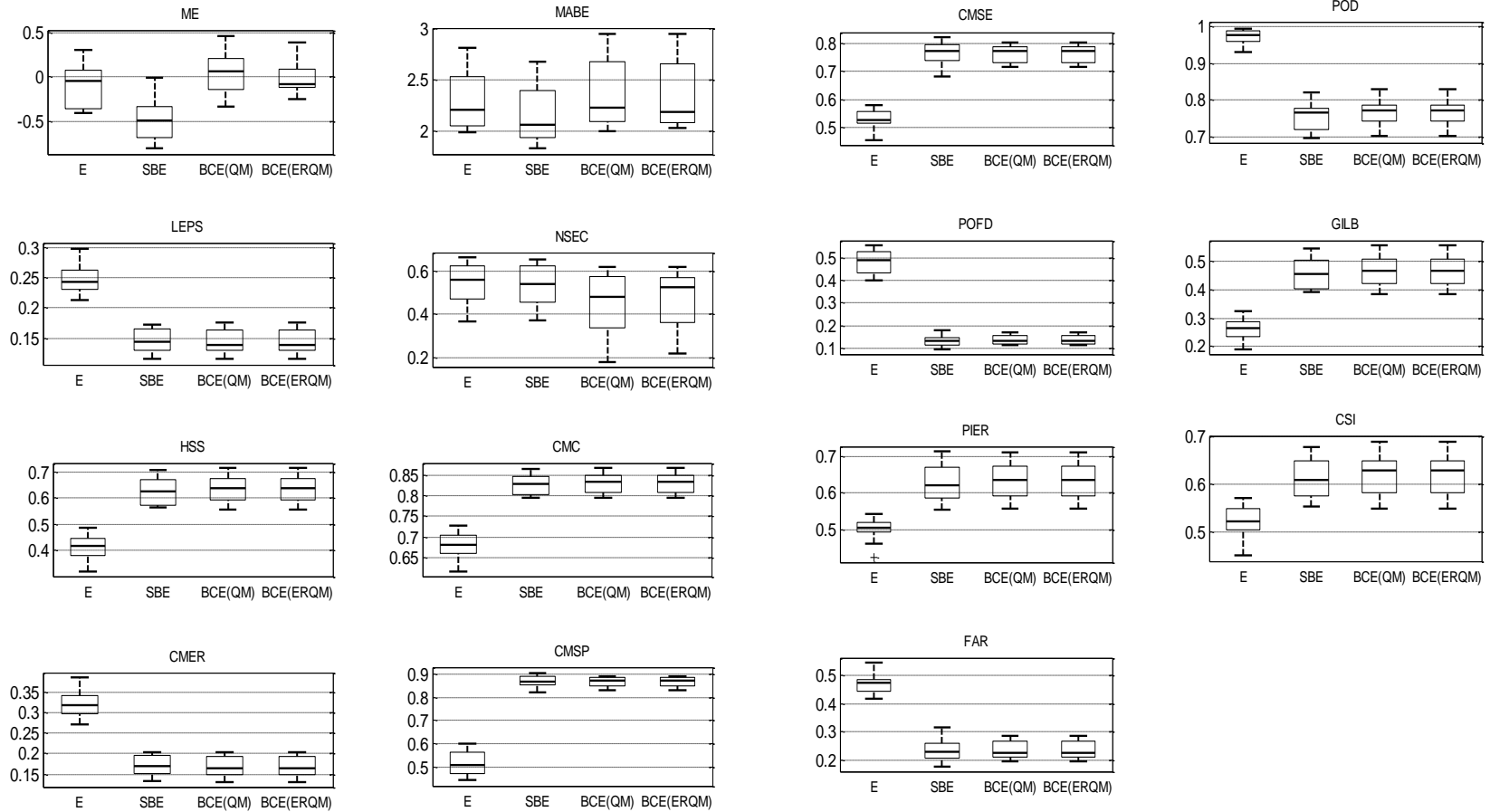
Observed and Estimated Quantiles



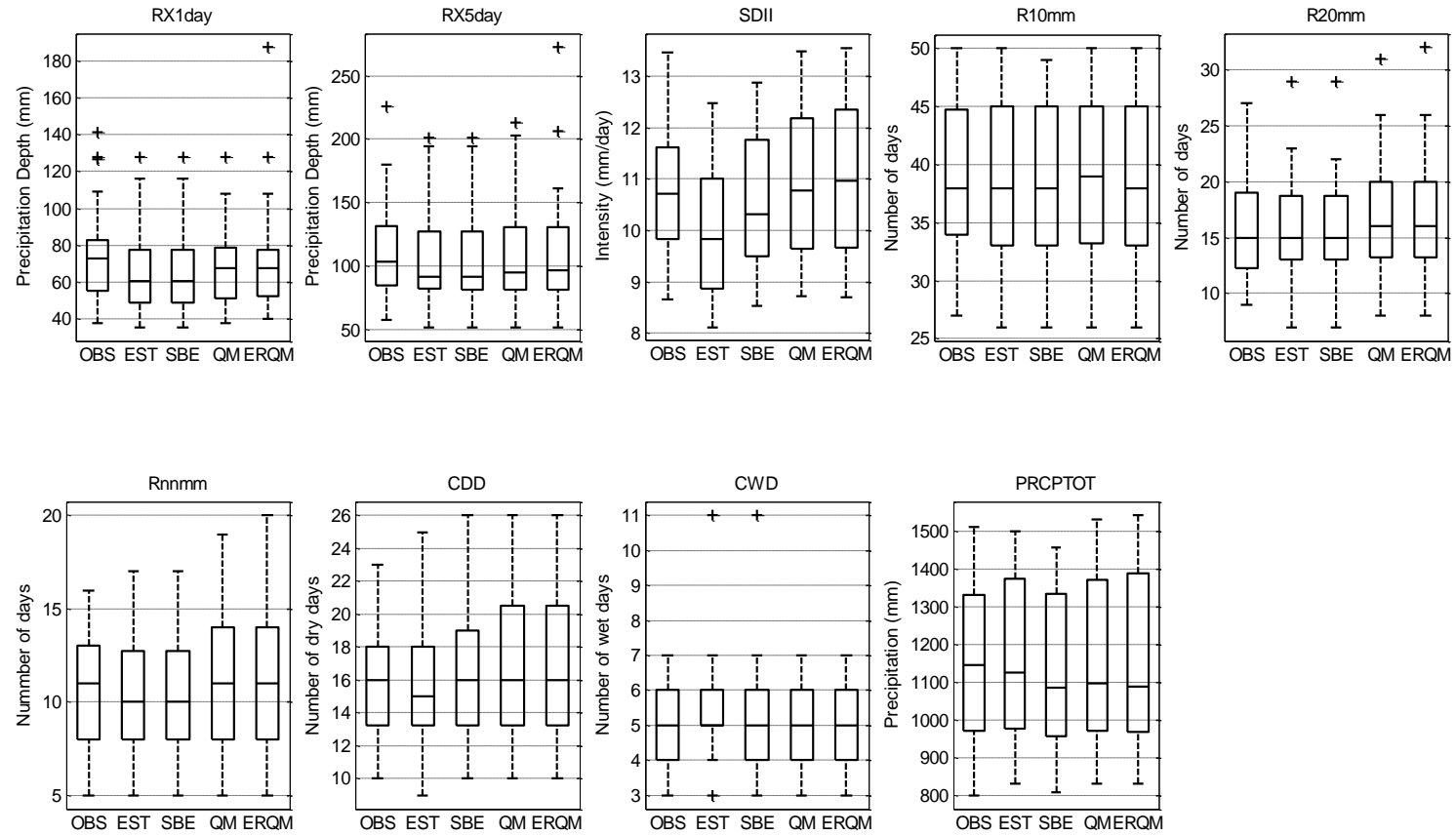
Post Correction Q-Q plots



Performance Measures



Changes in Precipitation Indices



-
- Comparative analysis of distributions of observed and filled test data can be carried out using nonparametric statistical hypothetical tests and they include:
 - 1) two-sample Kolmogorov-Smirnov (KS) (Smirnov, 1939; Sheskin, 2003);
 - 2) Ansari-Bradley (Ansari and Bradley, 1960) and 3) Wilcoxon rank-sum (Wilcoxon, 1945; Hollander and Wolfe, 1999). The null hypothesis (H_0) is that observed and estimated precipitation data are from the same continuous distribution. The alternative hypothesis (H_a) is that these two datasets are from different continuous distributions. The hypothesis tests are carried out at a statistical significance level of 5%.
 - The Ansari-Bradley (AB) test is used to evaluate the hypothesis (null hypothesis: H_0) that two independent samples of observed and filled data come from the same distribution against the alternative (alternative hypothesis: H_a) that they come from distributions that have the same median and shape but different dispersions (e.g., variances).
 - The rank-sum test can be used to evaluate the null hypothesis that observed and filled data are independent samples from identical continuous distributions with equal medians, against the alternative that they do not have equal medians. The test is equivalent to a Mann-Whitney **U**-test. The Wilcoxon rank-sum test is a nonparametric alternative to the t-test.

Observations

- Comparison of autocorrelation plots of observed and in-filled data series at different rain gauges indicate that serial autocorrelation values at different lags are underestimated.
- However, the underestimation is reduced when bias-corrected values are used. The autocorrelation plots of residuals based on observed and 1) spatially-estimated values and 2) bias-corrected values based on SBE, QM and ERQM for all rain gauges show that autocorrelation values at several lags lie within the 99% confidence bounds of the autocorrelation of a white noise process.

Observations

- This confirms that bias-corrections have not modified the error structure and preserved the independence of residuals.
- Bias-correction methods used in the current study attempt to match the CDFs of observed values from the training dataset and estimated precipitation and therefore the length and temporal window of training dataset will have implications on the performance measures based on corrections.
- Out of the three bias-correction methods evaluated in this study ERQM is most conceptually superior. SBE corrections focus only on dry-day revisions and the performance of QM is dependent on the stationarity of precipitation series.

-
- Statistical corrections of spatially-interpolated missing precipitation data estimates appear to be beneficial for improving estimates.
 - Missing precipitation data estimated at a site using an optimal spatial interpolation method are corrected using two different quantile-based bias-correction methods and an optimal single best estimator method.
 - Use of bias-correction methods eliminated the main limitation of the deterministic interpolation method in regards to over and under estimation of low and high extremes respectively at a site.

Spatial Analysis for Water Resources Modeling and Management

Methods for Analysis, Interpretation and Visualization of Spatial Data

Ramesh Teegavarapu, Ph.D., P.E.

Associate Professor,

Director, Hydrosystems Research Laboratory (HRL)

<http://hrl.fau.edu>

Department of Civil, Environmental and Geomatics Engineering,
Florida Atlantic University, Boca Raton, Florida, 33431, USA

Hydrometeorological Network Design

- Topics
 - Evaluation of point-based spatial observations,
 - Heterogeneity of the hydroclimatic processes represented by continuous surfaces.
 - Concepts of optimal spatial sampling schemes and variance-based methods for sampling network design.
 - Geostatistical approaches for monitoring network design.
 - Examples and applications of spatial analysis methods for optimal precipitation and solar radiation monitoring network designs.

Monitoring Network Design

- Design of (optimal) monitoring networks requires an evaluation of heterogeneity of the any variable based on observations at sampling locations in a region.
- Any optimal spatial sampling scheme requires a careful **balance** between sampling locations that are **too close to one another**, thus not providing enough **new information** (data highly auto-correlated), and sampling locations that are too sparse, so that processes at other spatial scales introduce too much variability

Approaches

- Several approaches exist in the context of any monitoring network design for a specific variable of interest, and they rely on conceptually simple methods based on variance of such variable in space.
- The variance of variable is calculated based on the existing number of observations (obtained spatially) in the region.

Available Methods

- A review of recent literature related to optimal design of monitoring networks points to several different currently available methods for a variety of applications and they include:
- Information-theoretic approach (Transportation applications (Xing *et al.* 2013) and wireless sensor networks (Larish&Riley 2011)).
- Entropy and multi-objective-based approach (Mogheir *et al.* 2013) and fuzzy theory and multiple criteria analysis (Chang & Lin 2014) for water quality monitoring.

Hydrometric Networks

- **A hydrometric network is aimed at giving the hydrological information to be used for the following needs :**
- Assessment of **the regional or national surface water resources** and of their trends (climatic and anthropogenic impacts)
- Water resources planning for management and utilisation
- Estimation of environmental, economic and social impacts of current or planned management practices on WR
- Analysis and forecasting of extreme events (warning) : drought, exceptional floods

Monitoring Networks

- Evaluation of existing networks
- Design of new networks
- Optimal re-design of existing networks
 - Removal of some stations
 - Relocation of some stations
 - Addition of some stations
 - All of the above.

Optimal Networks

- Optimal networks enable a good interpolation between the stations at any point in the area covered by the network, with enough accuracy for WRs management and utilization purposes.
- To optimise a network, is to find the best compromise between the richness and the interest of **hydrological information on the one hand**, and the **cost of acquisition of data on the other hand**

WMO* specified standard numbers

- Hydrometric Stations

Physiographic Units	Minimal Density per station <i>area in km² per station</i>
Costal zones	2750
Montaneous zones	1000
Interior plains	1875
Hilly Regions	1875
Small islands	300
Polar and arid zones	20 000

Surveys

- Survey among users of data on the usefulness of each station of the network ;
- – Multi-criteria analysis based on indicators which enable us to appreciate the usefulness of each station in the network and to characterize it

Methods

- Information-based methods
- Geostatistical Methods
- Multi-criterion-based methods

Sites based on trend

- Monitors that have a long historical record are valuable for tracking trends. In this analysis, sites are ranked based on the duration of the continuous measurement record.
- The analysis can be as simple as ranking the available monitors based on the length of the continuous sampling record.
- This technique places the most importance on
- sites with the longest continuous trend record.

Advantages

- Simple analysis, requiring few statistical tools
- Useful for identifying long-term trend sites
- A good first look at monitor history

Removal Bias

- Measured values are interpolated across the domain using the entire network.
- Sites are then systematically removed and the interpolation is repeated.
- The absolute difference between the concentration measured at a site and the concentration predicted by interpolation with the site removed is the site's removal bias.
- The greater the bias, the more important the site is for interpolation. This analysis can also be performed on groupings of sites to test various site removal scenarios.

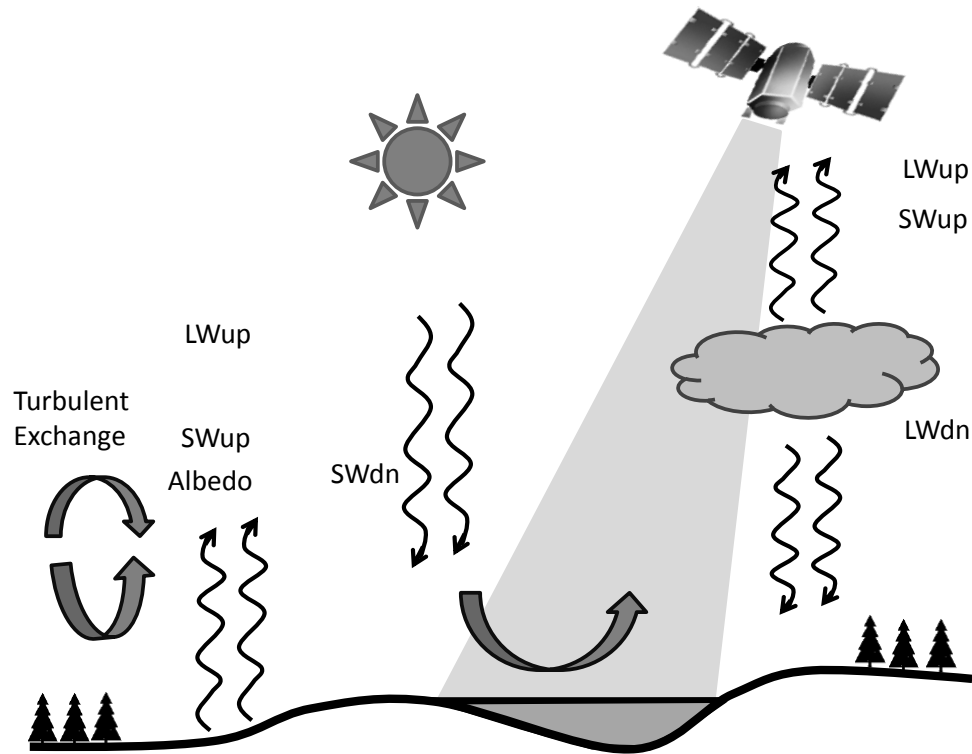
Example Network Design

- Solar Radiation Network Design

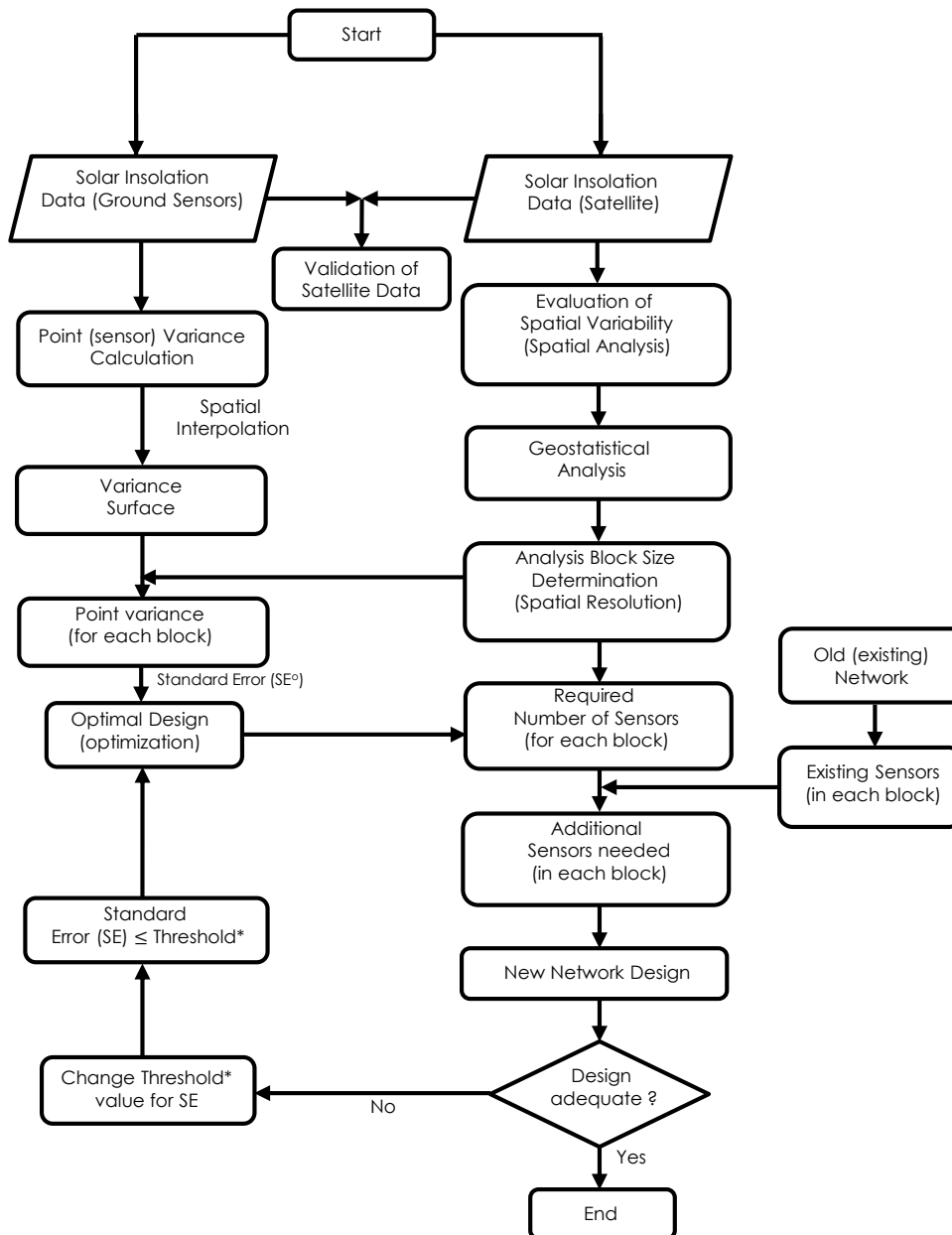
Example : Solar Radiation Network Design

- Design of a solar radiation network for ET estimation.
- ET can be estimated by several parameters including : temperature, solar radiation, etc.
- This particular work is based on solar radiation measurements that can be used for estimation of ET.
- Solar radiation is measured using pyranometers at different locations.

Solar Radiation



Schematic of simplified energy balance model. Here, “up” and “dn” relate to the upward and downward components, respectively. “SW” is shortwave radiation, and “LW” is long wave radiation.

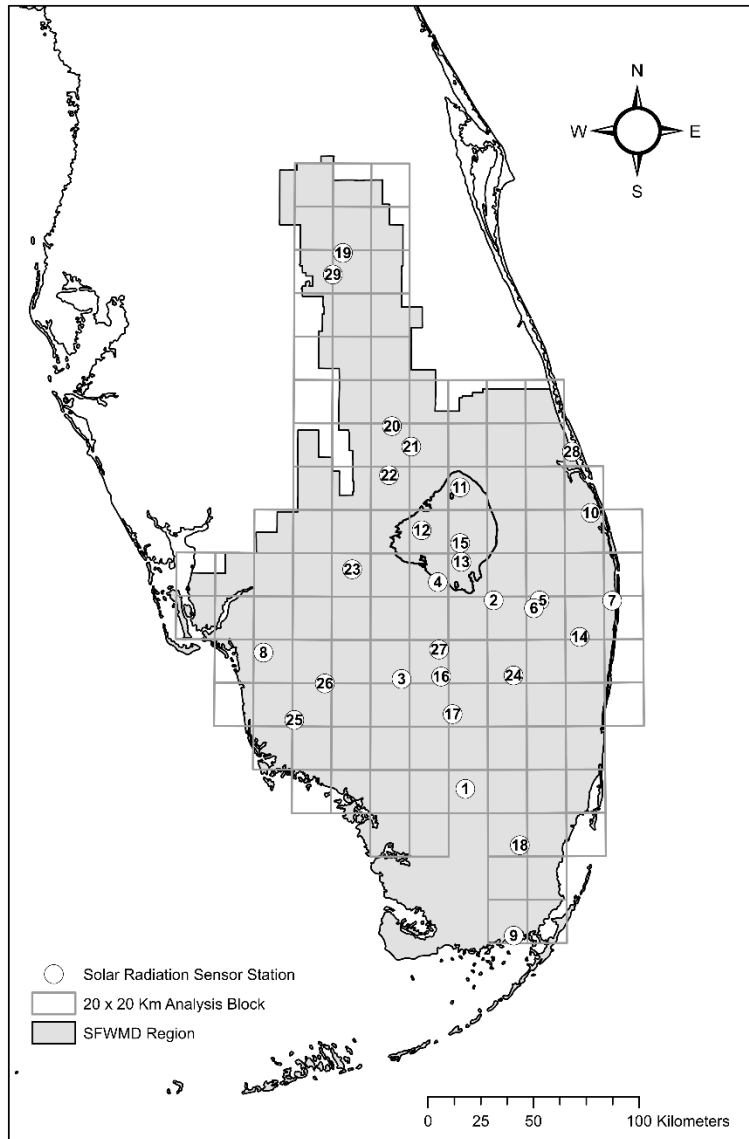


Initial value of SE: SE°

Step followed in the design of monitoring network for solar radiation

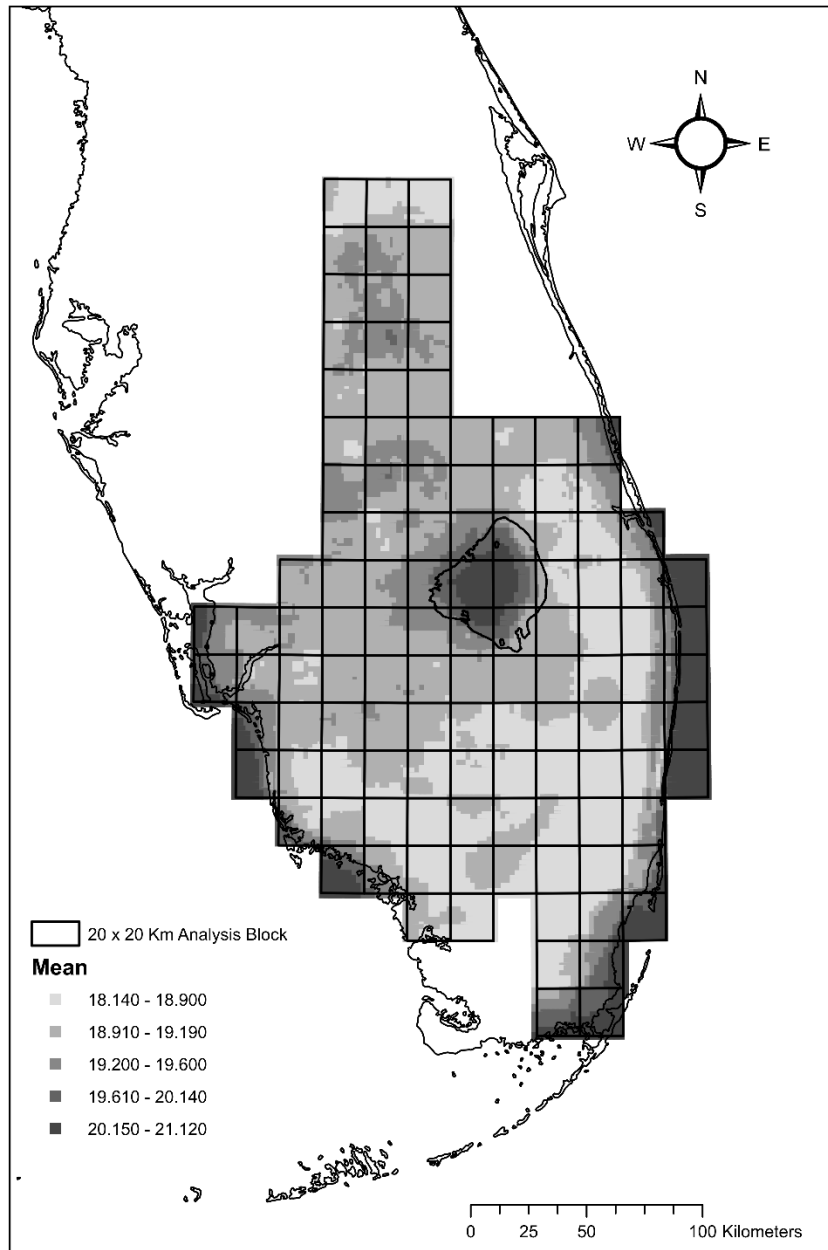
The methodology uses observed solar radiation data from solar radiation sensors and also satellite-based gridded data.

Geostatistical approach is used for design.



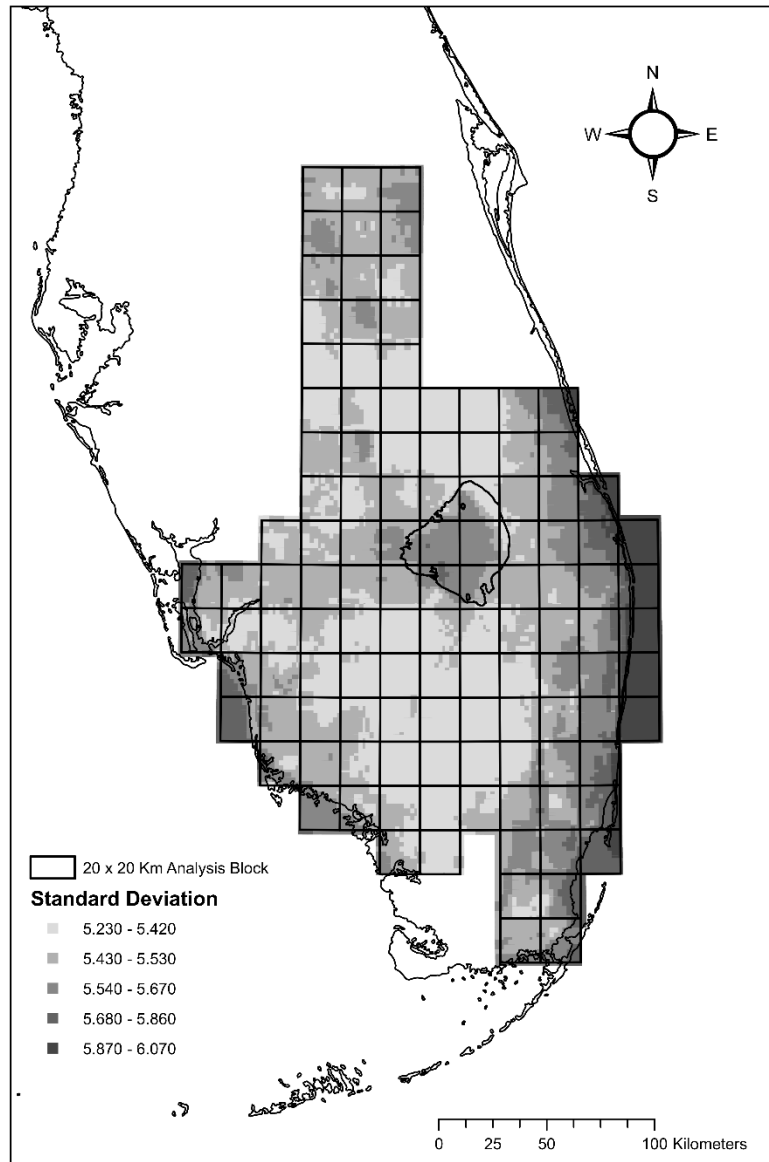
Location of solar radiation sensors (observation stations).

These sensors measure solar radiation and are referred to as pyranometers

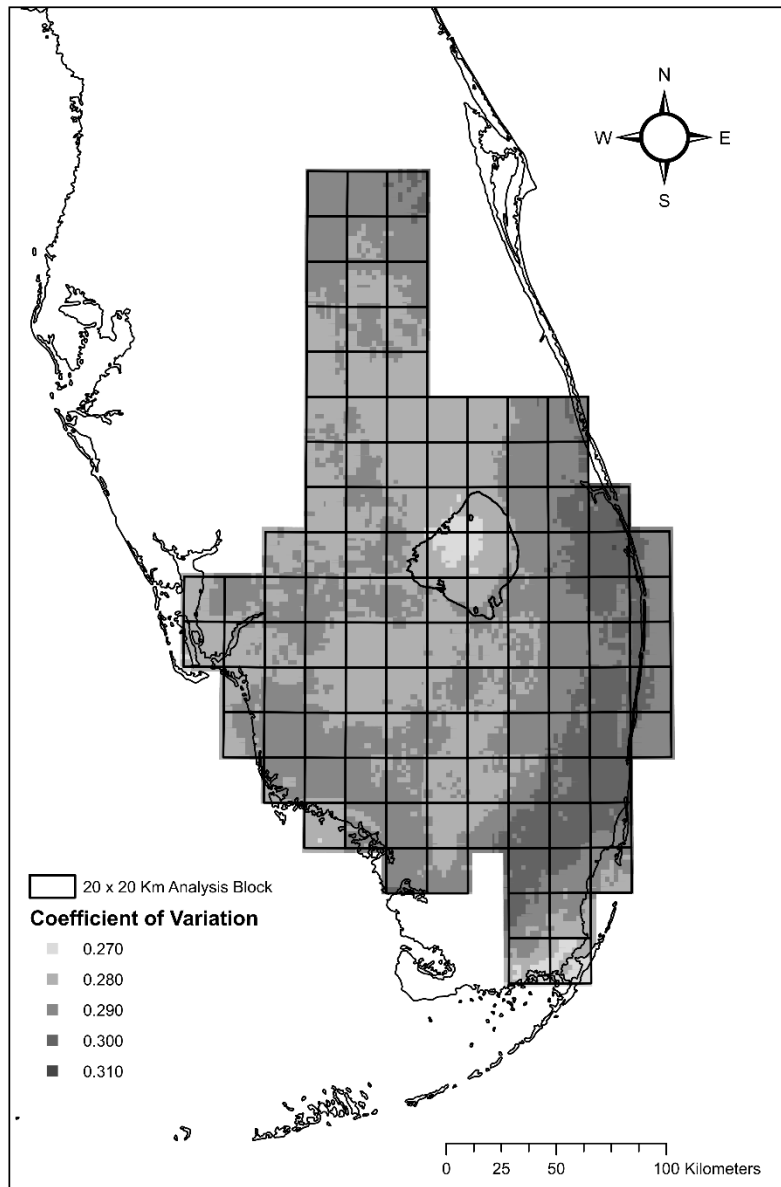


To understand the spatial variability of the radiation, satellite based gridded data is used and is analyzed.

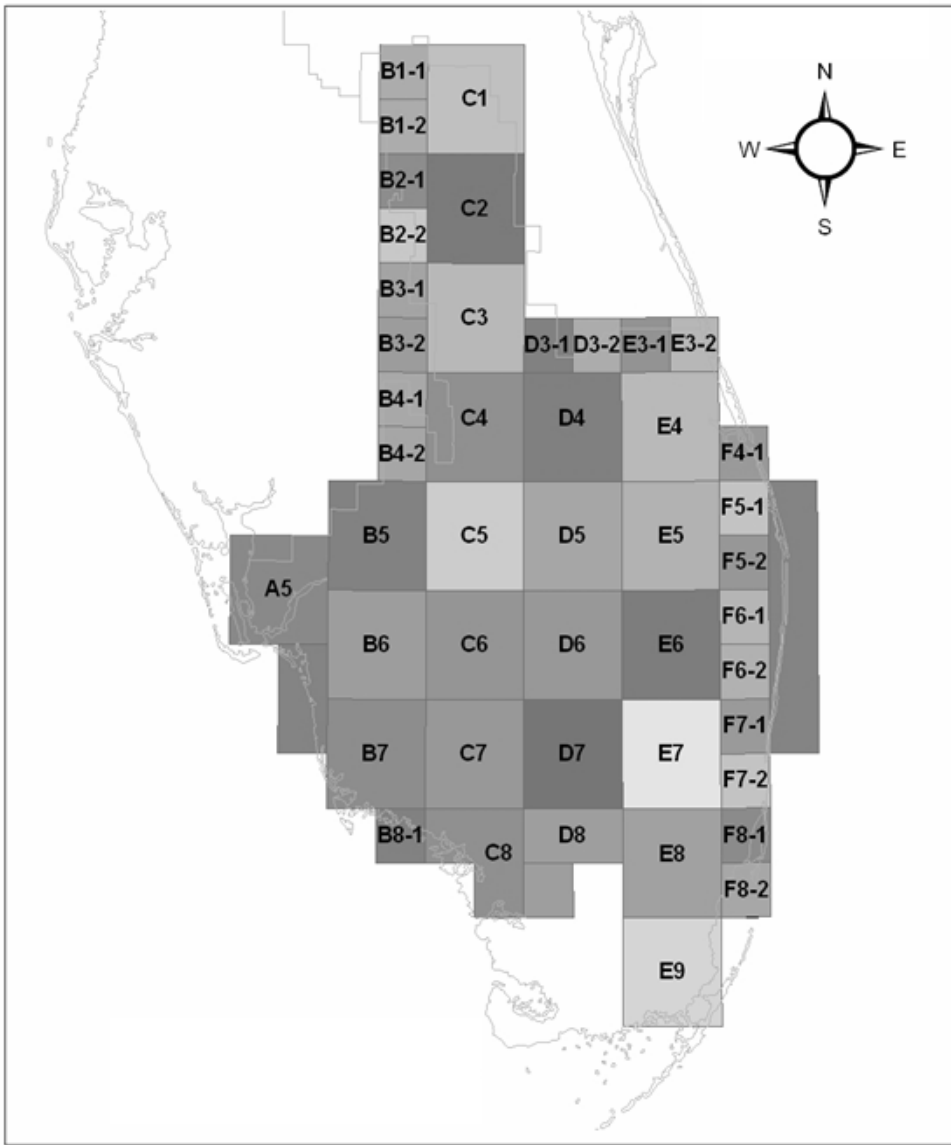
Mean solar radiation values shown in the map indicate higher values in the coastal areas as well as high values in the lake (located in the center of the region)



Standard deviation values of solar radiation shown in the map indicate higher values in the coastal areas as well as high values in the lake (located in the center of the region)



Coefficient of variation values of solar radiation shown in the map indicate higher values in the coastal areas as well as high values in the lake (located in the center of the region)



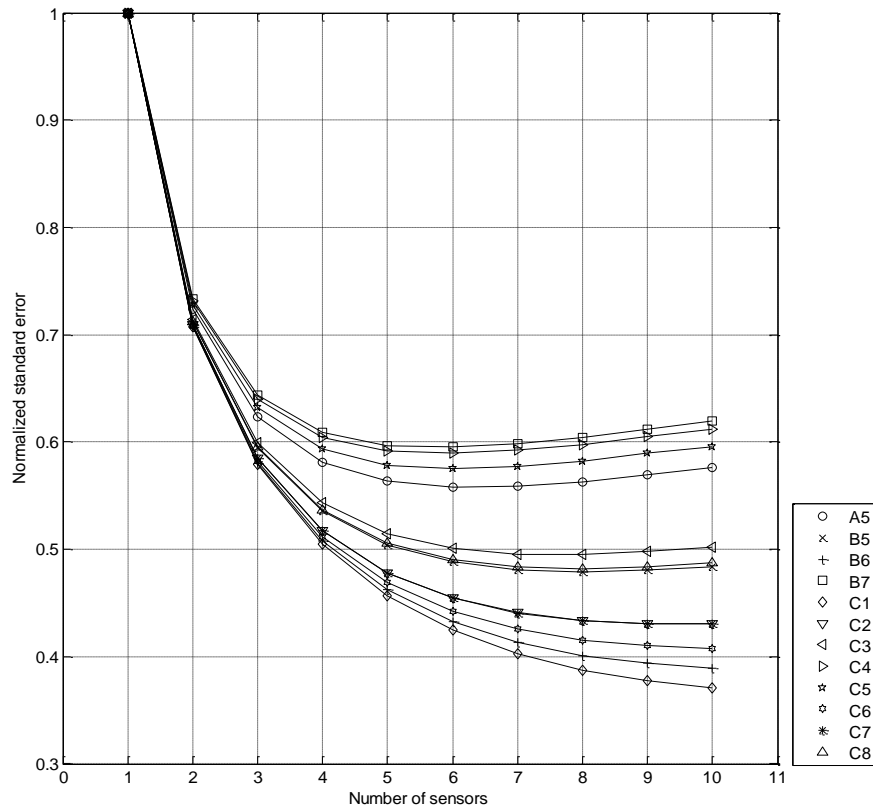
Based on the variability of the radiation, analysis blocks are defined. These blocks are defined based on the existing grid size of 2 km x 2 km. A varying analysis grid size is used for the region, with finer grid resolution in areas with high variability and coarser resolution in areas with low variability.

Please see the published paper for more details included with the notes.

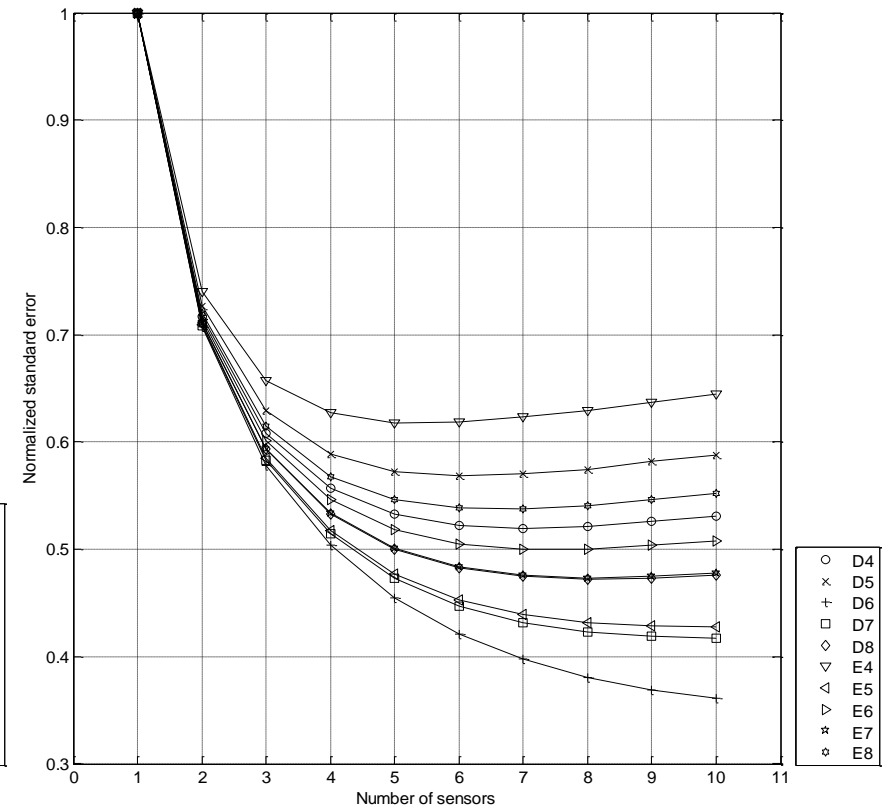
Design

- The design of the sensor network is aimed at capturing and characterizing the spatial variability of the solar radiation across the region.
- The design of the network depends on several factors, including:
 - (1) placement of sensors to maximize the information obtained from the sensors,
 - (2) the existing network of sensors and
 - (3) the monetary cost involved in the purchase and placement of the sensors.
- The standard error (SE) of the mean is used as a metric of accuracy of the monitoring network, to identify the optimal number of ground sensors in this study

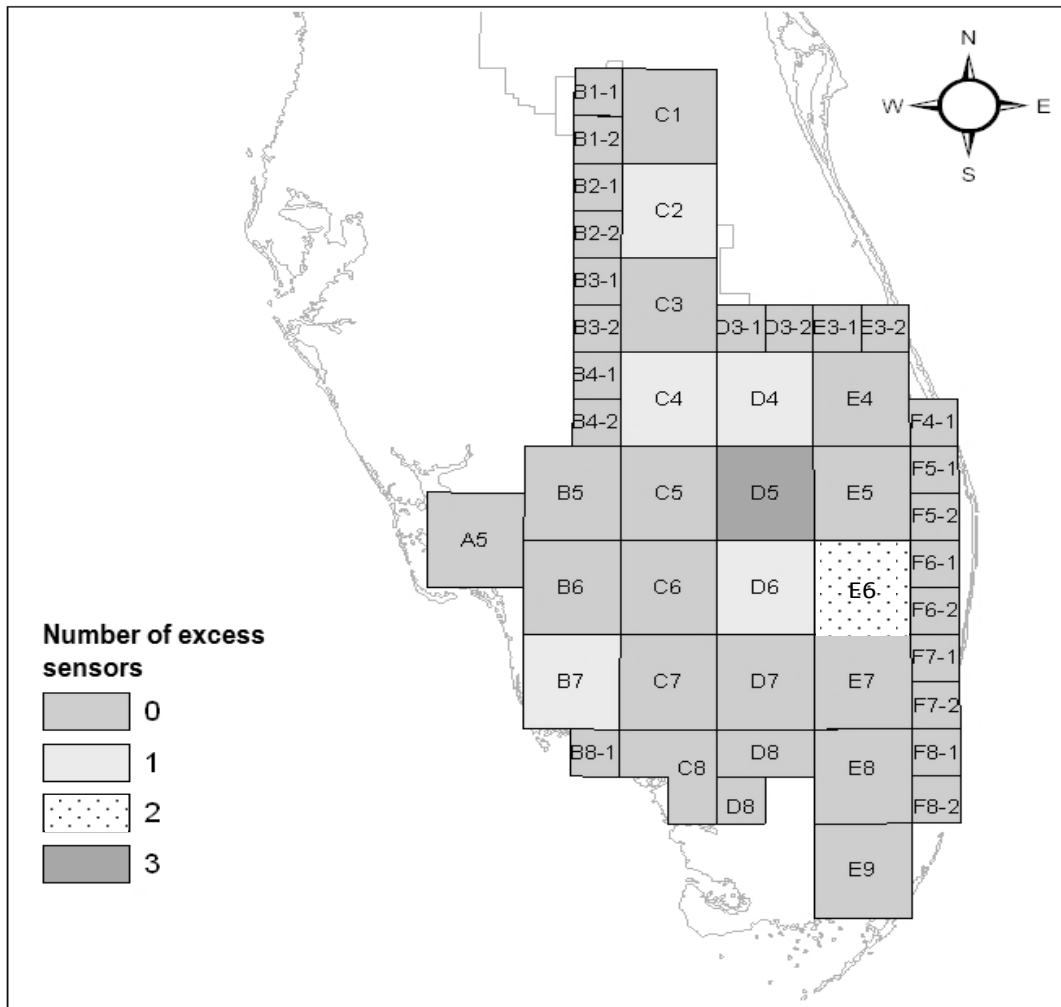
Standard Errors



(A)

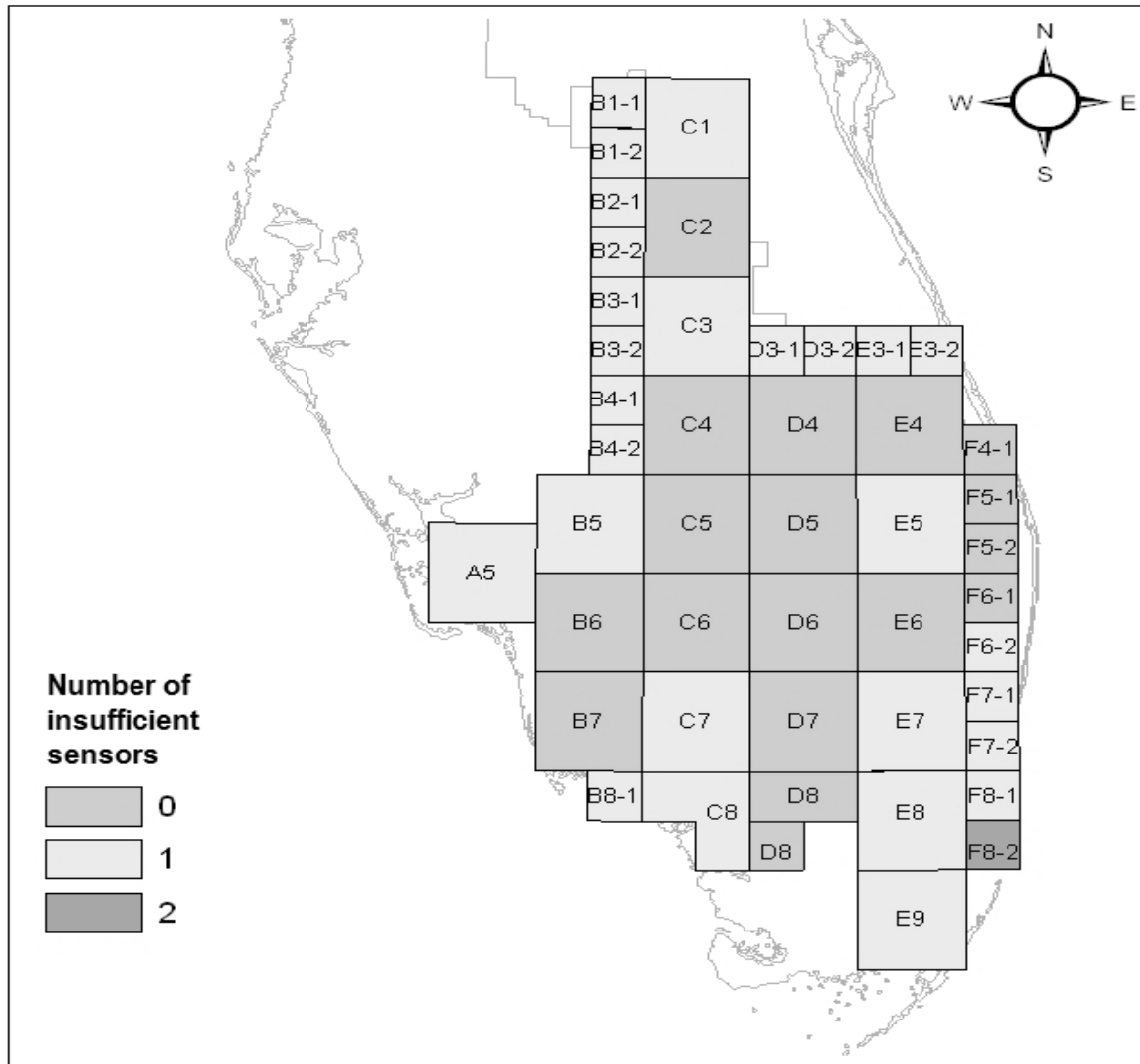


(B)



For each analysis block the optimum number of sensors is obtained.

Once this number is obtained, information about the existing (already existing) sensors is used to obtain the number of excess sensors



For each analysis block the number of insufficient sensors is obtained.

Once this number is obtained, additional sensors required for each block is recommended.

Number of sensors required and SE

Table 6. Variations in the number of sensors required and additional needed (new) for different values of standard error.

Standard error (MJ day ⁻¹ m ⁻²)	Number of sensors		
	Required	Available	New
3.5	124	29	95
3.8	97	29	68
4.5	79	29	50
5.0	48	29	19
5.5	45	29	16

Table 7. Number of sensors required in different analysis blocks for a specific standard error.

Standard error (MJ day ⁻¹ m ⁻²)						Standard error (MJ day ⁻¹ m ⁻²)					
Block	3.5	3.8	4.5	5	5.5	Block	3.5	3.8	4.5	5	5.5
A5	2	2	1	1	1	B1-1	2	2	1	1	1
B5	2	2	1	1	1	B1-2	2	2	1	1	1
B6	2	2	1	1	1	B2-1	2	2	2	1	1
B7	2	2	1	1	1	B2-2	3	2	2	1	1
C1	2	2	1	1	1	B3-1	3	2	2	1	1
C2	2	2	1	1	1	B3-2	3	2	2	1	1
C3	2	2	1	1	1	B4-1	3	2	2	1	1
C4	2	2	1	1	1	B4-2	3	2	2	1	1
C5	2	2	1	1	1	B8-1	3	2	2	1	1
C6	3	2	2	1	1	D3-1	3	2	2	1	1
C7	3	2	2	1	1	D3-2	3	2	2	1	1
C8	3	2	2	1	1	E3-1	3	2	2	1	1
D4	3	2	2	1	1	E3-2	3	2	2	1	1
D5	3	2	2	1	1	F4-1	3	2	2	1	1
D6	3	2	2	1	1	F5-1	3	2	2	1	1
D7	3	2	2	1	1	F5-2	3	2	2	1	1
D8	3	2	2	1	1	F6-1	3	2	2	1	1
E4	3	3	2	1	1	F6-2	3	2	2	1	1
E5	3	3	2	1	1	F7-1	3	2	2	1	1
E6	3	3	2	1	1	F7-2	3	2	2	1	1
E7	3	3	2	1	1	F8-1	3	2	2	1	1
E8	3	3	2	2	1	F8-2	3	3	2	2	1
E9	4	3	2	2	1	Total	63	45	42	23	22
Total	61	52	37	25	23						

Observations

- A methodology to design an optimal sensor network to characterize the spatial variability of solar insolation useful for estimation of evapotranspiration (ET) in a region is proposed and evaluated in this study.
- Application of the methodology to upgrade an existing network of solar radiation sensors (i.e. pyranometers) in a region of South Florida, USA, is reported. Geostationary operational environmental satellite (GOES) satellite and ground sensor network-based data are used in this study for the design of the network.

-
- The optimal network is expected to improve the estimation of ET using a simple solar-radiation-based ET estimation method in the region.
 - An array of analysis blocks with two different fixed spatial resolutions (20 and 40km) is defined based on the evaluation of spatial variability of solar insolation data in the study region.

-
- Geospatial and geostatistical analyses are used to assess the solar insolation within each analysis block and to obtain an optimal number of sensors.
 - Results from the analyses conducted in this study indicate that the number of sensors required in each analysis block depends on the standard error (SE) set as a criterion for network measurement accuracy.

-
- An optimal sensor network that is expected to provide a standard error (SE) of $5.0 \text{ MJ day}^{-1} \text{ m}^{-2}$ is selected to demonstrate the utility of the proposed methodology in this study.
 - A separate study needs to be carried out to clearly define the implementation strategies for the recommended network developed in this study

-
- The methodology proposed and evaluated in this study is generic and can be used for design of an optimal monitoring network for any hydro-meteorological variable.

Additional Material

- Additional notes, papers, Matlab codes are included along with the material.

-
- Thank you.

Dr. T. RameshTeegavarapu Ph.D., P.E.

Associate Professor, <http://faculty.eng.fau.edu/ramesh>
Director, Hydrosystems Research Laboratory (HRL), <http://hrl.fau.edu>
Department of Civil Environmental and Geomatics Engineering
Room 217, Building # 36
777 Glades Road, Boca Raton, Florida, 33431
email: rteegava@fau.edu, ramesh.teegavarapu@gmail.com
Phone: 561.297.3444
Fax: 561.297.0493.