Geography 4203 / 5203

GIS Modeling

Class 11: Sampling and Core Areas

Some Updates - Projects!

- Projects!!!
- 12 March deadline for **data demo**, ready for use…
- Make sure you have the data at hand which you intend to use during the project
- *21 March*: Class devoted to **group meetings**

  - groups discuss the further **proceeding** and have access to Stefan and Jeremy before Spring break
- After the Spring break (after 31 March): some classes are devoted to **discussions** of project-related issues if needed

Some Updates - Progress Reports

- 14 & 16 April **Progress reports** (presentations, ca. 15 min) -- should cover two class meetings

  - problems, state (where are you?), focused points

  - **tasks**, responsibilities for each **group member** (not everybody has to present)

  - Getting **advices** / **help** from the class

  - **Discussions**

Some Updates - 1-2-1 Meetings

- 18 & 21 April **One-to-One meetings with project leaders**

  - Getting direct feedback about the group work, how you organize the group and how you judge the progress the group made so far to solve the problem

  - Are there any problems, difficulties to organize the work or to define working packages

Some Updates - Final Presentations

- 23, 25, 28 & 30 April **Final presentations (20-25min + discussion points)**

  - **coordinated** by the project leader

  - **reiteration** of the problem statement,

  - **summarizing** what was (or was not) accomplished

  - **all group members** participate in the verbal presentations

  - describing their **specific role** and describing **results** and what they learned in the final presentation

Some Updates - Final Report

- 10-12 pages text double-spaced, absolute maximum) including tables, charts and maps and a **problem description** (something like in the proposal)

- **methods** used to solve the problem, any **problems** the group found (data cleaning problems, projection issues, etc.) and how the group finally did **solve** them

- **project leaders**: short essay discussing **project management** (tasks delegation, team work: in parallel on different or the same task, or on tasks as a group) + What would each group leader have done differently if they had the group project to do all over again?

Some Updates - Grading

- **Group level (25 points)**:

  based on the difficulty of the project, the technical execution, on the clarity / logic of class presentations, and quality of the written report

  Group dynamic (participation, involvement)

- **Individual (25 points)**:

  based on the work each individual has been done on the project (progress report, final report & presentation, observation)

- **Project leader (20 points)**:

  project design and management (goal orientation, involvement according to skill levels between members, contact for help)

Last Lecture

- We had a look at **dasymetric mapping** and **areal interpolation**
- You have seen some interesting examples how dasymetric principles and methods can be used to **improve** our data
- You understood how dasymetric mapping makes use of **ancillary data** to **disaggregate** aggregated data (often in choropleth maps)
- You took home the message that you can improve your data by making use of **ancillary** data that more **reflect** the **underlying statistical surface** and have a **relationship** to our variable of interest

Today‘s Outline

- We will use the time to repeat and deepen **sampling** issues
- What are the pros and cons for different sampling designs in relation to the analysis we are performing and the **area** and its **variation** in attribute values
- Some more methods for spatial estimation such as **core area** mapping (e.g., convex hulls, kernel mapping) which are related to what we have seen so far

Learning Objectives

- You will better understand different **sampling strategies** and you will know how to evaluate them (you already know how to judge the usability of samplings for your analysis)
- You will learn alternative **methods** where you can make use of **estimation** principles in spatial analysis
- You will see some **examples**

Sampling Basics

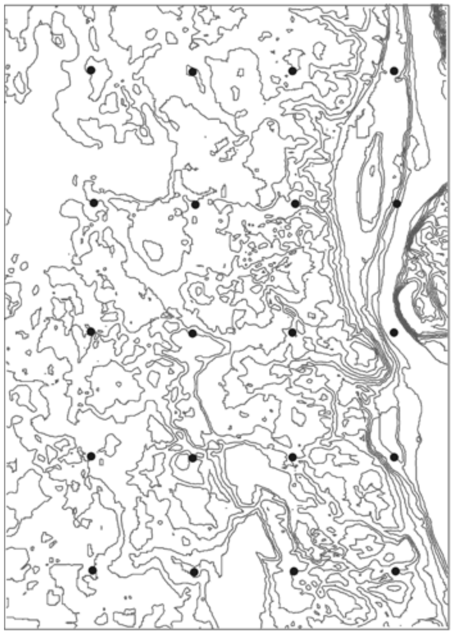
- Aim of estimation is to find values for a variable at **unknown locations** based on values measured at **sampled** locations
- Planning important to make the sampling more efficient / accurate
- **Control** taken over locations of sample points (**patterns/dispersion**) and **sample size**
- Sometimes neither can be controlled (**diseases** within a population)

Control of Sample Size

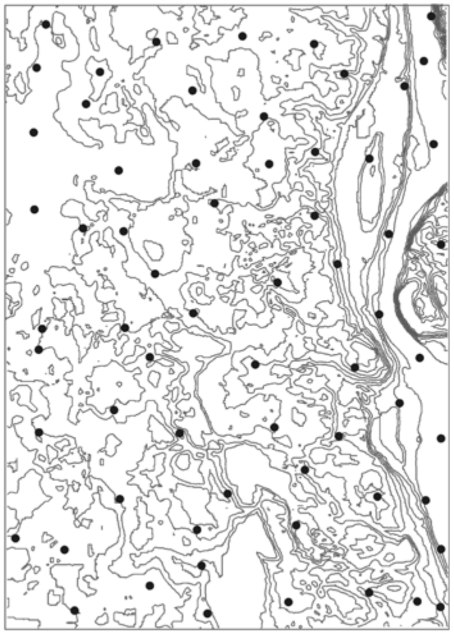
- Law of **diminishing returns**: Situation where further sample points add relatively little **additional information** or **gain** in **accuracy** for substantially **increased costs**
- The rule is: most surfaces from interpolation are **undersampled** (funds as limiting factor)
- Difficult to determine the optimal sample size for interpolation methods

Control of Sampling Patterns

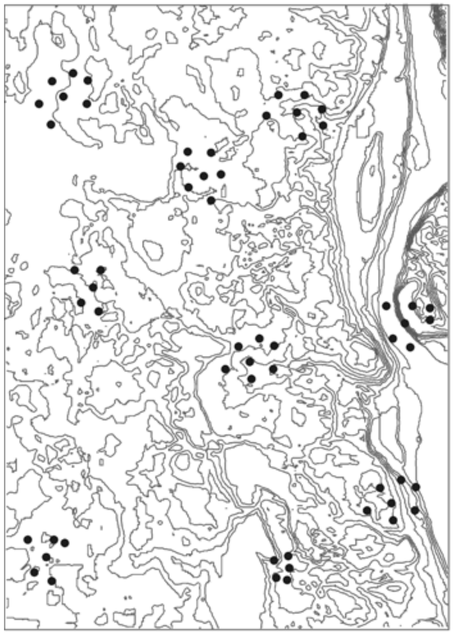
- Sample locations **spread** across our working area - so it's important **how to collect** sample point data
- **Patterns** we choose affect the quality of the interpolation carried out (and have effect on **sample sizes** needed)
- Wrong **distributions** increase estimation errors and simply cost money…
- **Systematic**, **random**, **cluster**, **adaptive/ stratified, transect and contour** sampling
- Remember sampling from photogrammetric data sources: regular, progressive, selective, composite



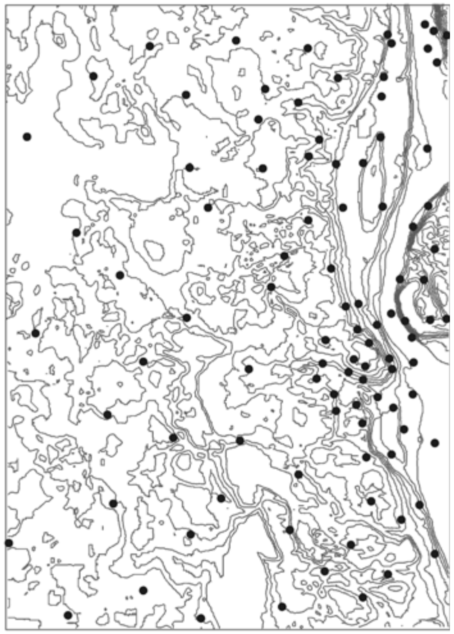
a)



b)



c)



d)

Control of Sampling Patterns: Systematic Sampling I

- Sample points spaced **uniformly** at fixed x- and y- intervals (not necessarily the same in both directions) along **parallel** lines
- x and y axes not required to align with the **northing** and **easting** grid directions
- Advantage: **Ease** in **planning** and **description**, little **subjective** judgement

Control of Sampling Patterns: Systematic Sampling II

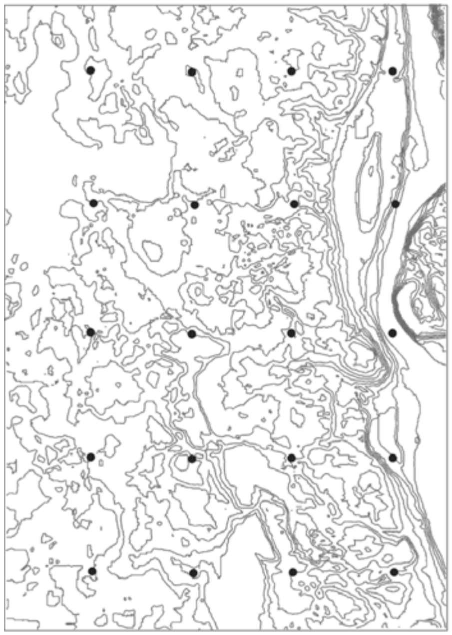
- Disadvantages:

  - Reduced **statistical efficiency** (equal sampling **intensity** in all areas and preferences cannot be addressed)

  - **Accessibility** to the sample points (no difficult terrain addressed) and **travel costs**

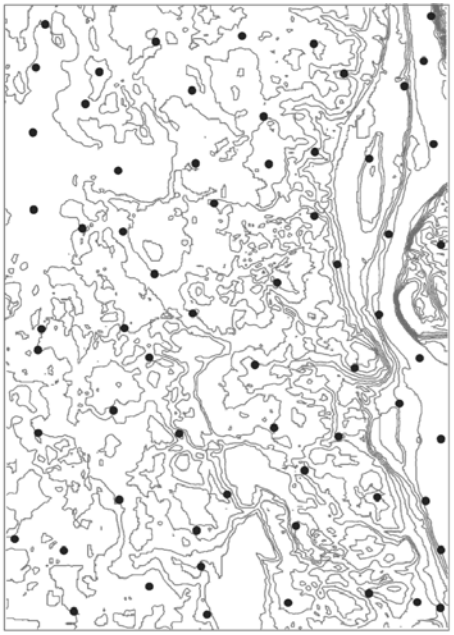
  - Potentially **biased** estimations

  - **Oversampling** of **overproportioned** areas can result in interpolation error of other locations



Control of Sampling Patterns: Random Sampling I

- Avoids some of the **disadvantages** of systematic sampling
- Point locations based on **random number generation** (easting and northing independent random numbers)
- **GPS measurements** to find the locations
- **Low risk of bias** and inaccurate estimations



Control of Sampling Patterns: Random Sampling II

- Improvement only **limited**
- Areas of **higher variation** are not necessarily sampled more intense
- More points collected than necessary in **uniform areas**
- Fewer points than needed in **variable areas**

Control of Sampling Patterns: Cluster Sampling I

- **Grouping** of sample points based on defined criteria
- Clusters around **cluster centers** such that distances between points **within** one cluster are smaller than the distances **between** cluster centers
- Advantage: **reduced travel** time (cluster points close together), important for **natural resource** surveys
- Problem: **Variation not considered** explicitly

Control of Sampling Patterns: Cluster Sampling II

- Establishment:

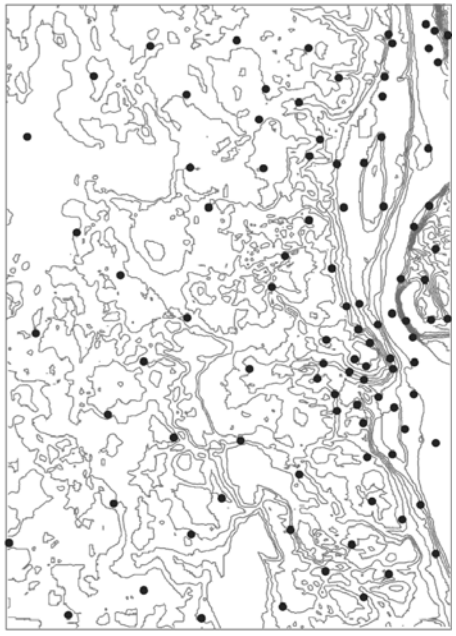
  centers located **randomly** or **systematically**

  sample points of each cluster may also be placed **randomly** or **systematically** around the center

  U.S. **forest survey**

Control of Sampling Patterns: Adaptive Sampling I

- **Stratified** sampling
- **Higher sampling density** where the feature of interest is more **variable**
- Increased **sampling efficiency**
- **Small-scale variation** addressed
- **Homogeneous** areas are represented by few sample points whereas areas with **higher spatial variation** are sampled with higher density



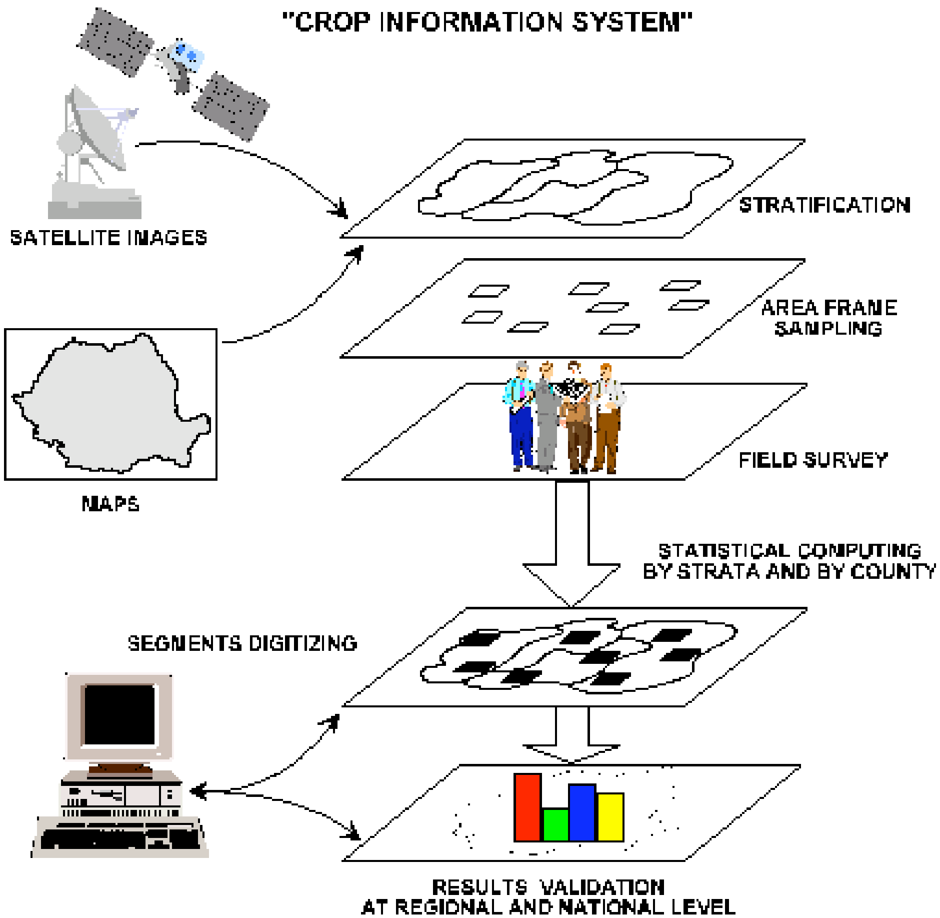
Control of Sampling Patterns: Adaptive Sampling II

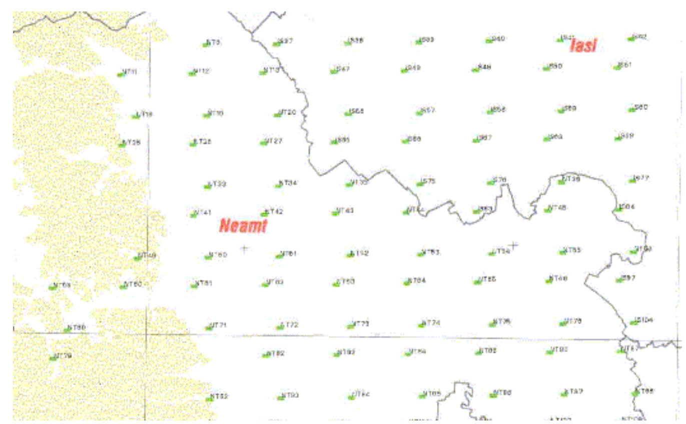
- **Preliminary estimation** of **feature variation** in the field is required
- For example: based on **steepness** observations in the field **sample density** can be defined for individual **strata**
- If variation cannot be observed on the spot **preliminary maps** are produced and **local variation** determined for a **repeated field visit**

Some More Things About Sampling Design

- **Area frames**
- Total populations
- **Remote sensing** data sources
- Sampling **without replacement** for area objects etc.
- **Multi-phase** sampling, **multi-stage** sampling

Pilot Study Crop Information Systems in Romania (FAO)





Sampling and Vicinity

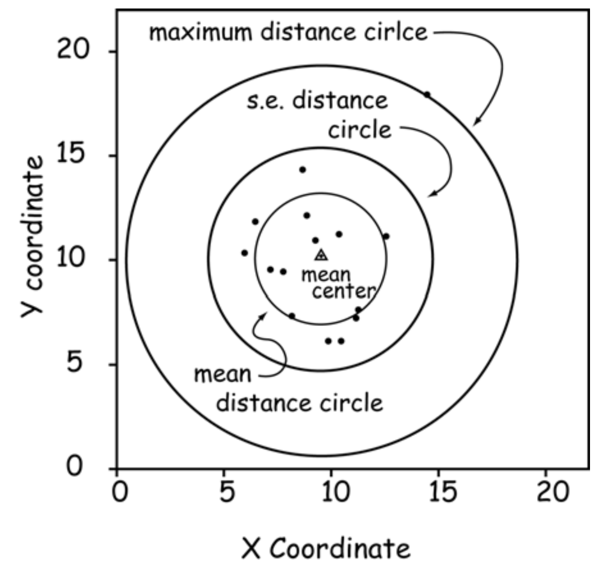
- What do you know about the relationships between these two terms?
- Where is vicinity explicitly unwanted?
- Where are we implying vicinity and why and how and what for…?
- What do think about before defining a sample across our area?

Core Area Mapping

- Another useful spatial analysis tool
- Core Area is a **primary area of influence** or activity for the feature of interest
- Area that characterizes **high values** for a variable / event (high use, density, probability of occurrence,…)
- Crime analysis
- Habitat analysis for species
- Customer identification in space

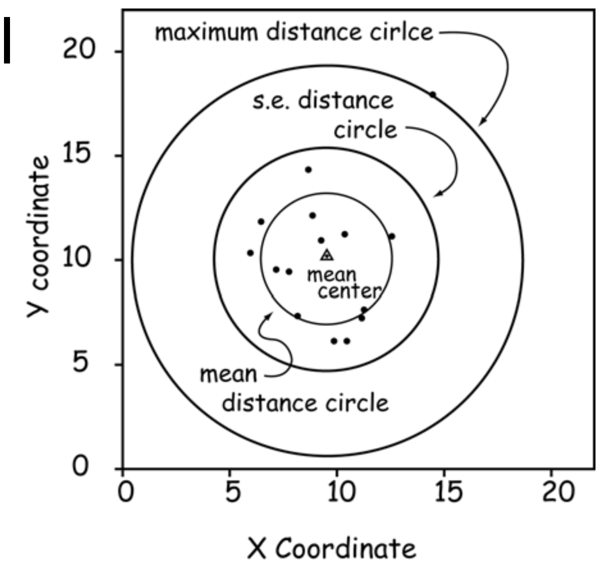
Concept of the Core Area

- Variable often as sample **points** or **line** observations
- We are interested in identifying the **area features** derived from this set of observations
- Core area is a **higher dimensional** spatial object than the underlying observations
- Finding the area where searched features **occur frequently**
- For example: centers of criminal activity, endangered species and key habitat conditions identified



Mean Center & Mean Circle

- **Mean center**: Average x and y coordinates of the sample points
- **Mean circles**: Mean center as central point and a radius defined based on some statistical measures (max distance, average distance, variance of distance values)
- Simplistic in approach but assume **circular shape** of the area and thus are potentially biased by **outliers**
- Also: reality is mostly **irregularly shaped**

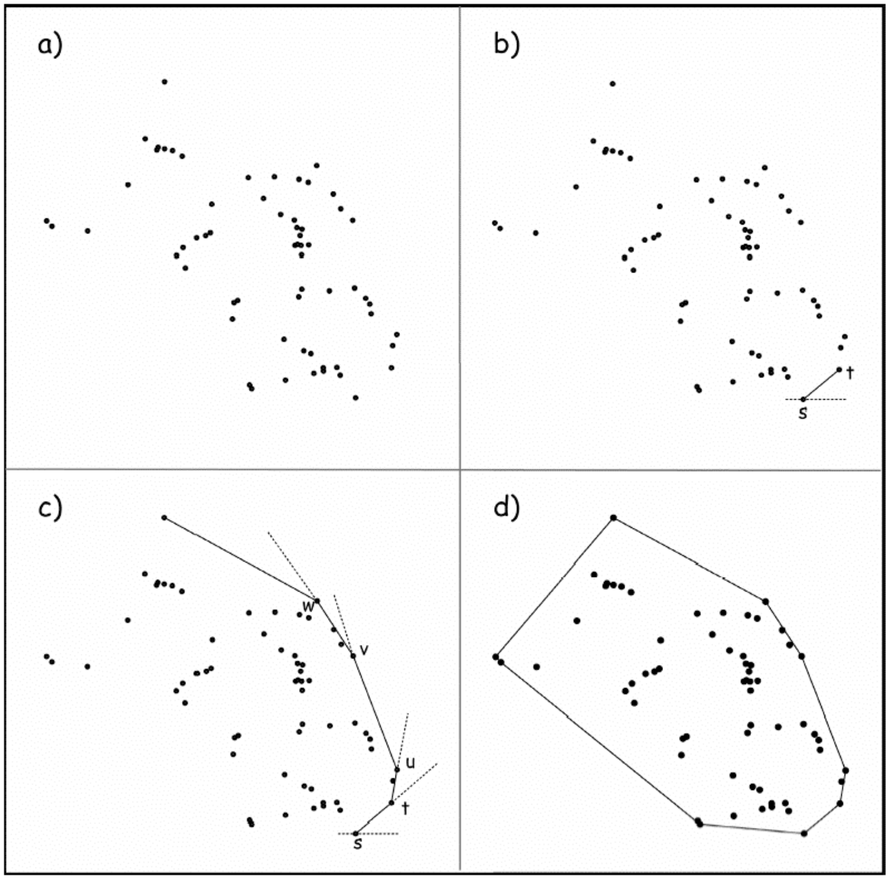


Convex Hulls

- “Minimum convex polygons” to identify core areas with irregular shapes
- Def: The smallest polygon created by edges that **completely enclose** a set of points and for which all exterior angles between edges are **greater or equal 180 degrees**

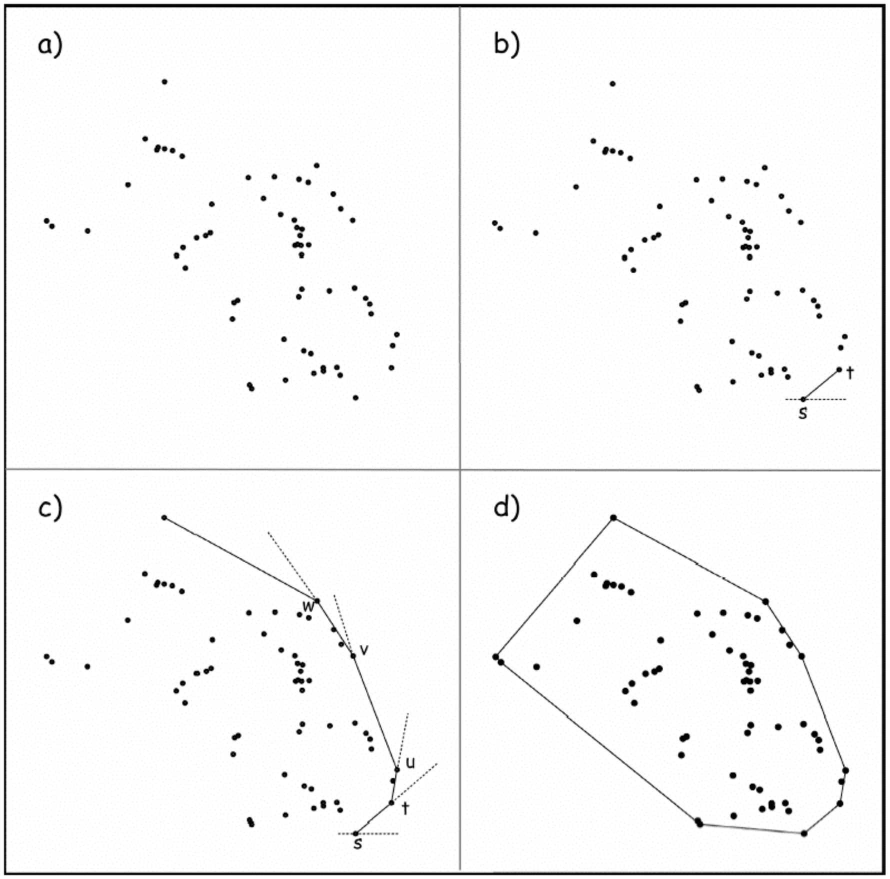
Convex Hull as Natural Bounding

- Reflects the **irregular shape** based on a set of points
- But: sensitive to **outliers** -- the CH becomes **unreasonably large**
- Low level of **subjectivity**
- Pure **spatial arrangement**
- **Sweep algorithm**



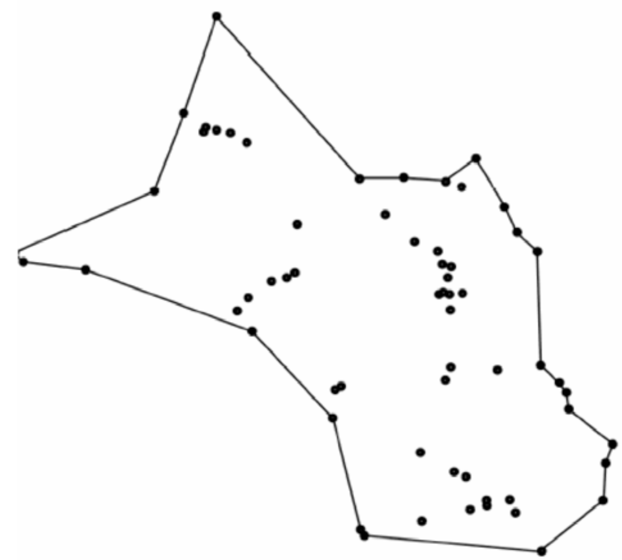
Sweep algorithm

- "**Extreme**" point identified
- Angle of **deflection** from this one to **all other** points determined
- Point with the **smallest** positive clock-wise or counter-clock-wise **angle** identified
- This one is the next point of the **CH** and is used for the **next calculation**
- Process **repeated** until **starting point** is reached



Convex Hulls -- Some Issues

- Ignore **clustering** in the data
- Potential **loss of information** on **density** and **frequency** of occurrence in the **interior** region
- Algorithms for deriving **concave polygons** have been developed but rarely used



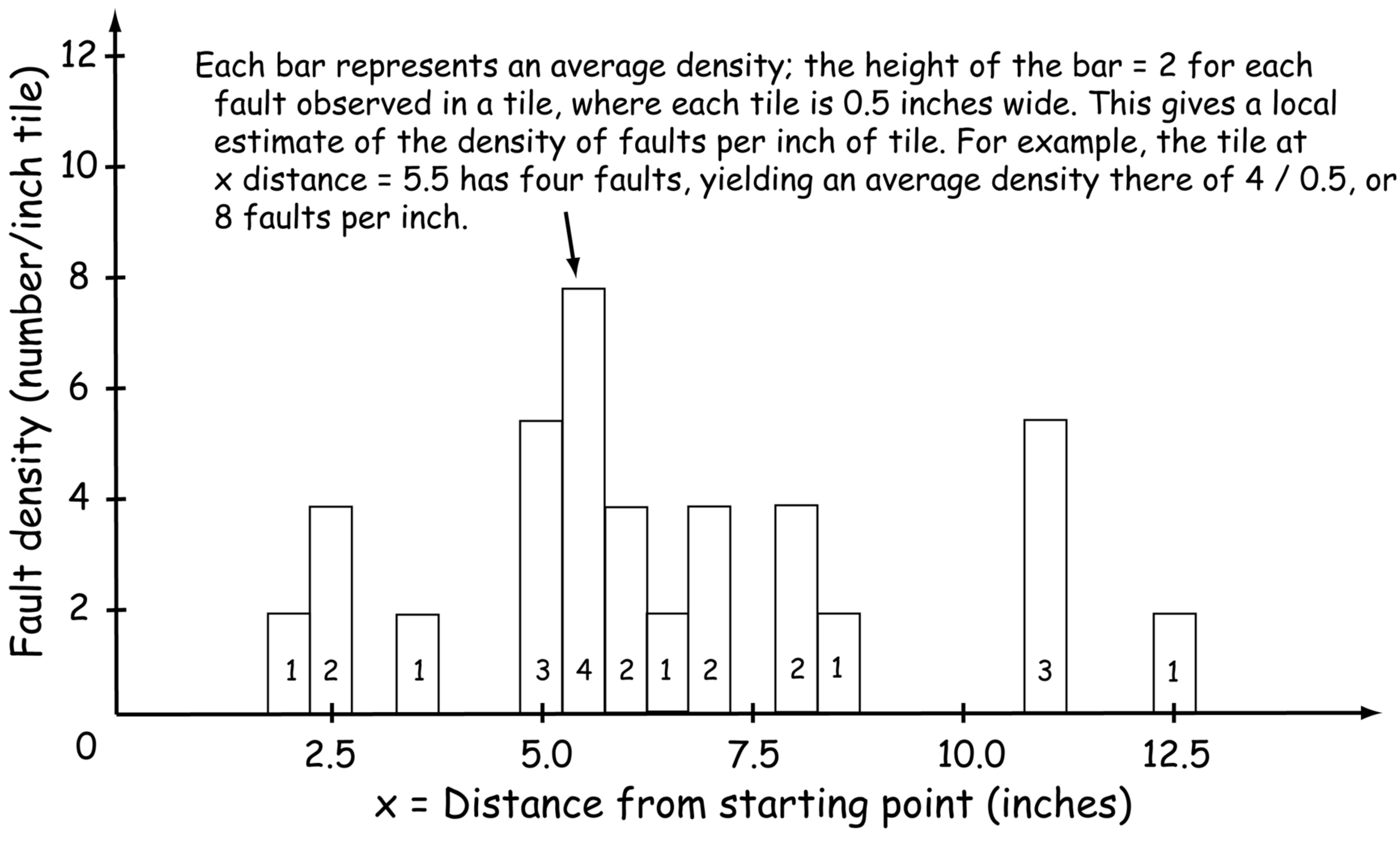
Kernel Mapping

- Using a set of sample locations to estimate a **continuous density surface**
- Mathematically **flexible**
- Ease of **implementation**
- Robust to **outliers** and considers **clustering**
- Can represent **irregular shapes** and can be **statistically-based**

Kernel Mapping - The basic Idea

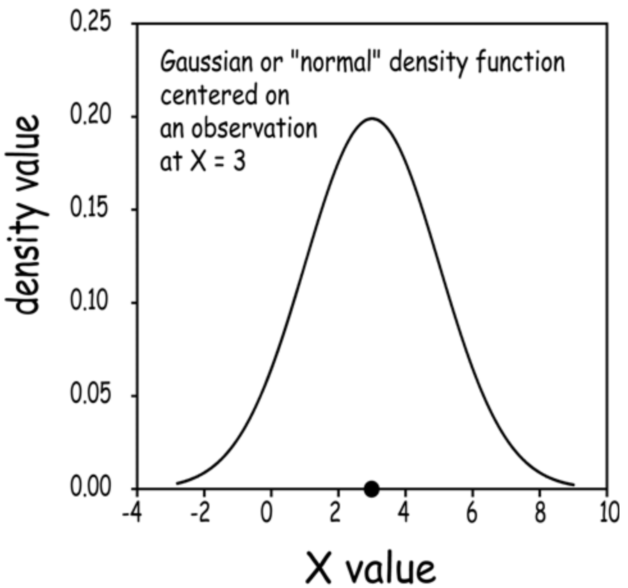
- Based on **density distributions** assumed for each sample point
- These density distributions are placed over the **same plane** (one for each point) and **vertically added (cumulative)**
- Result is the **Composite Density** of our sample
- This Composite Density can be used for **Core Area Mapping** (densest areas first)

Local Densities - Along a Line Example



Local Densities - Some Issues

- Assuming a **characteristic shape** for the density function (for **each point**)
- Rectangle in the tile example (**uniform** density across the tile)
- In reality rather **symmetric** shapes (parabolas, Gaussian curves,…) to produce **smoothly varying surfaces**
- Shapes can be **mathematically** defined



Local Density Functions

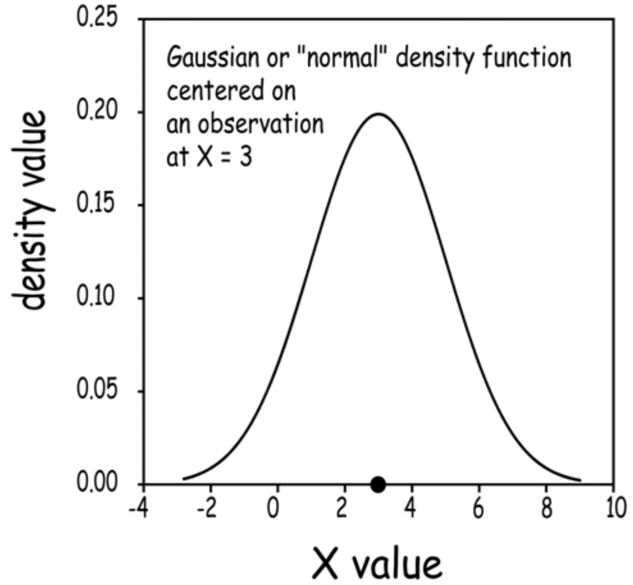
- E.g., **Gaussian** curve as a symmetric function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

  μ - sample location

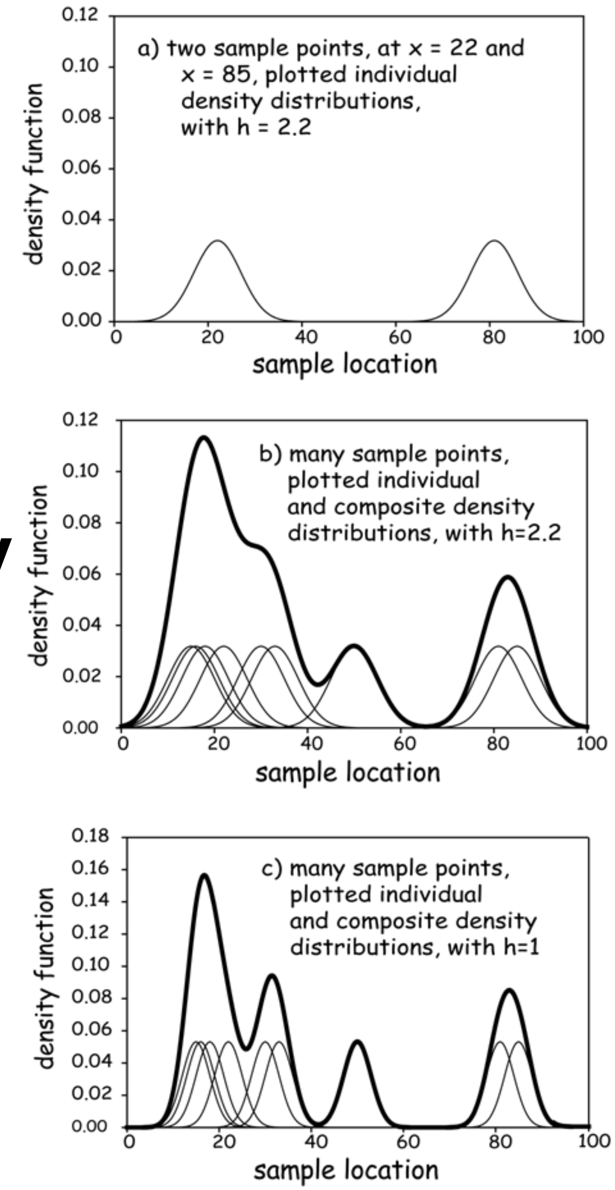
  σ - scaling constant

- **Peak** at μ; **area** = 1
- Shape: how fast is the **peak** reached, how **pointed** is the peak, how quickly are values **returned** to **zero** at distant points???

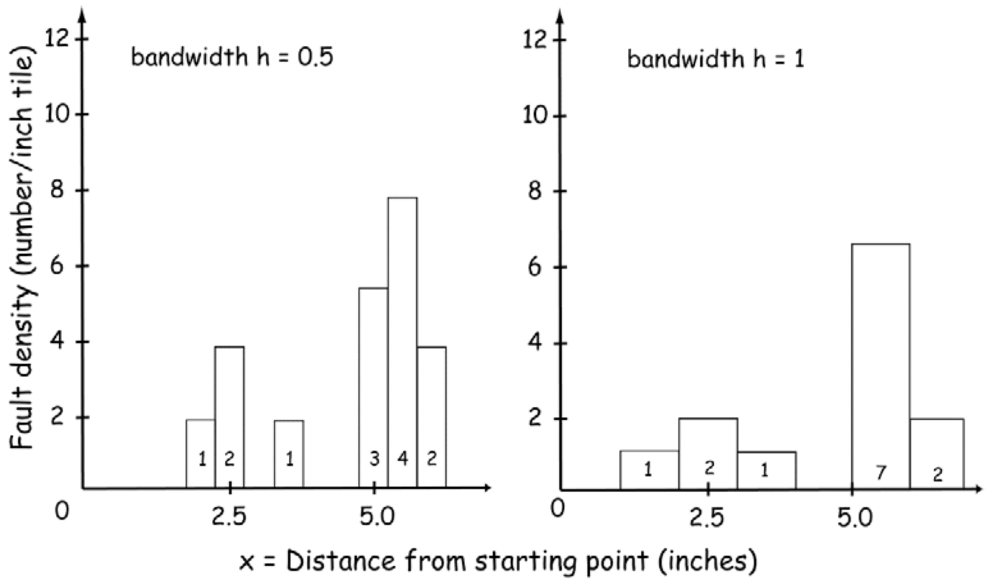


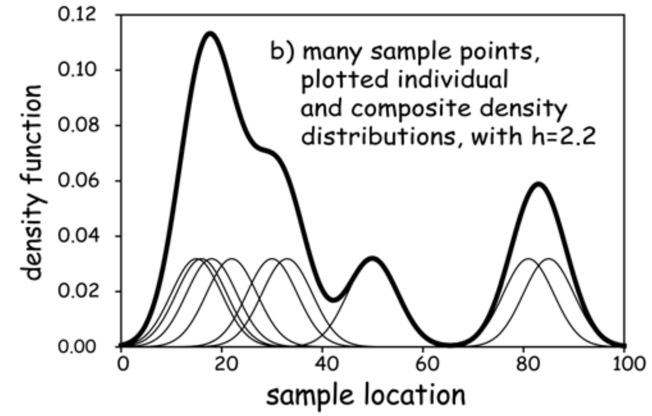
Composite Density Distribution

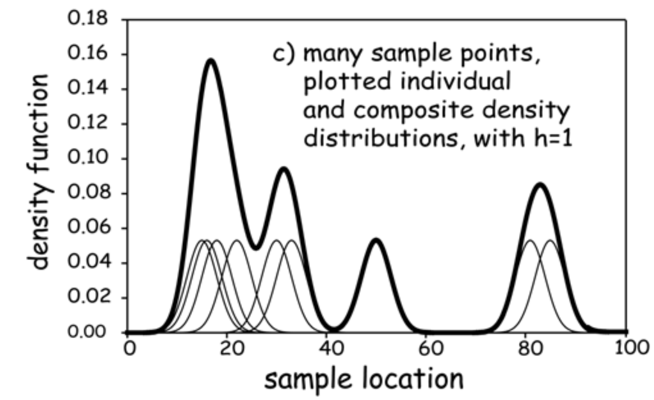
- "**Stacking** the individual density distributions
- **Overlapping** bumps if all points are plotted are summed **vertically**
- **Cumulative** density distribution
- **Bandwidth** or "**spread**" of the individual density distributions (**binning** intervals)



Bandwidth Effects







Steps in Kernel Mapping

- **Collection of sample points** (locations and attributes)
- Choice for one **Kernel density function**
- Define **bandwidth h**, apply the **Kernel density distribution** and sum **composite density** estimates for each location

$$\lambda(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n}\frac{K(x_i,y_i)}{h}$$

- $\lambda(x,y)$ -- comp. density distribution
- n -- number of samples
- h -- bandwidth
- K -- ind. Density distribution (applied at each i)

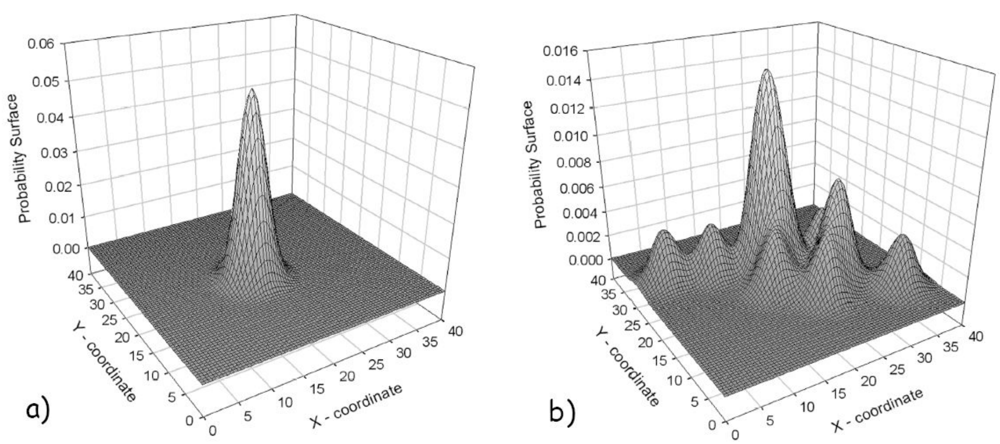
Presentation of Density Distributions

- **Density** or **probability** of occurrence of the underlying variable

  crime density mapped within a city

  defect density in a tile floor

  density of plant species within a home range



Definition of the Bandwidth h

- Difficult to determine the best value of **h**
- Commonly, we would apply **several density surfaces** for different **h** and look at the result to find the best **approximation**
- By observing the change of density values over a range of values h getting **insight to the data**
- Formulas exist and have very different outcomes:

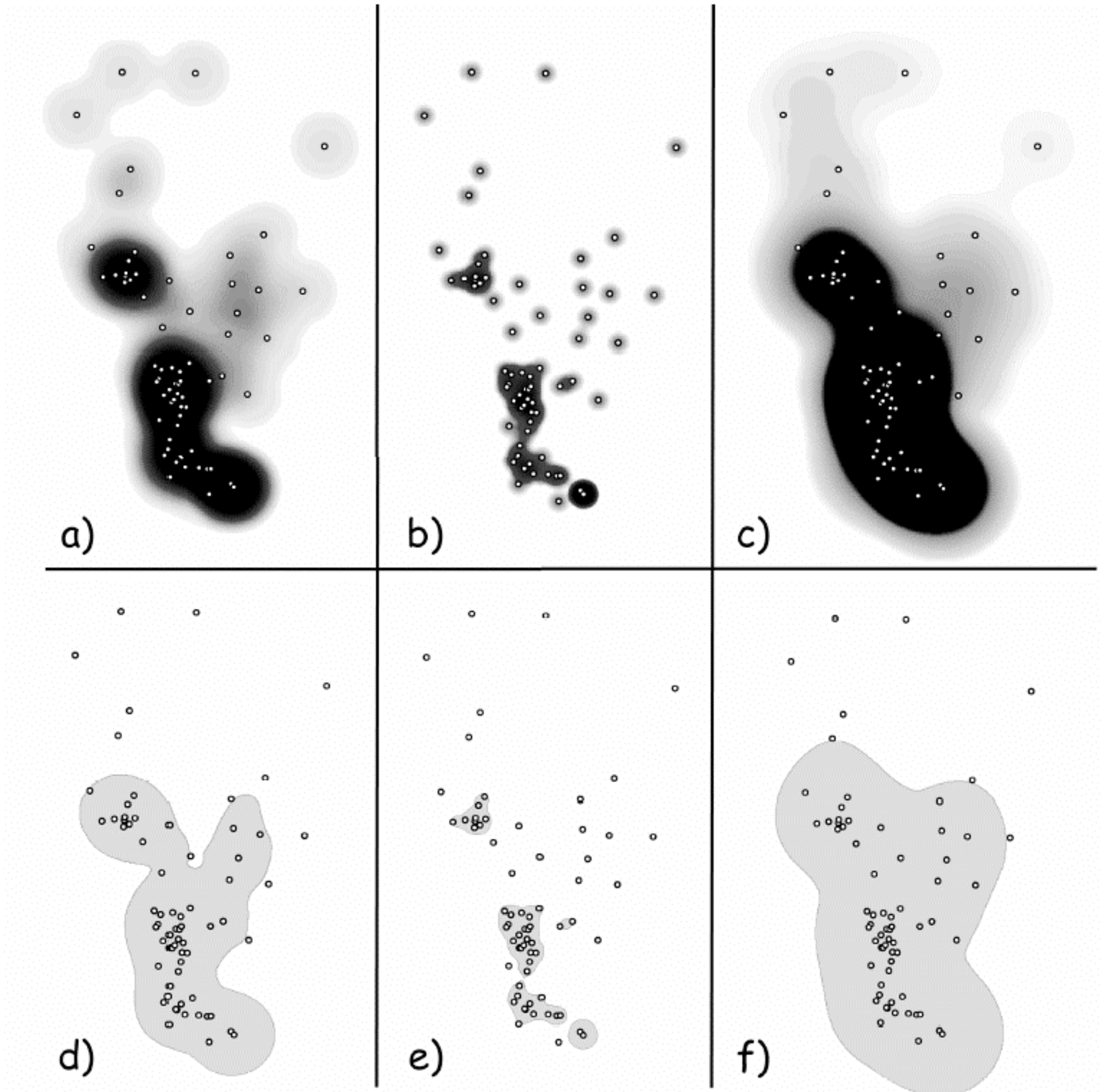
$$h_{opt} = \left[\frac{2}{3n}\right]^{\frac{1}{4}} \sigma$$

h_{opt} -- optimal bandwidth
n -- number of samples
σ -- standard deviation (unknown but estimated from the sample)

Fotheringham et al. 2000 for a Gasussian Kernel

Core Area Delineation

- Derived from the density surfaces by defining **value thresholds**
- Dependent on the **bandwidth** chosen
- Different band widths result in different **core area polygons**



optimum Below optimum Above optimum

for values of > 90% in the density surface

Summary

- Core area mapping implies some very interesting approaches to **convert point data** into **higher-dimensional** spatial data
- We have looked at **mean center**, **convex hull** and **Kernel mapping** approaches
- Kernel mapping creates **continuous** surfaces of **density** or **probability** of the underlying variable
- Depending on the **bandwidth** chosen and the the **local density** distribution the **delineated core area** polygons can be different
- Also depending on **thresholds** defined…