KIG-C1010 Introduction to geoinformatics

Lecture 4: Introduction to vector data analysis

Aalto University
School of Engineering

Jussi Nikander

20.1.2023

Topics for today

- Map overlays and joins
- Spatial data normalization
- Density surfaces and interpolation
- Network analysis

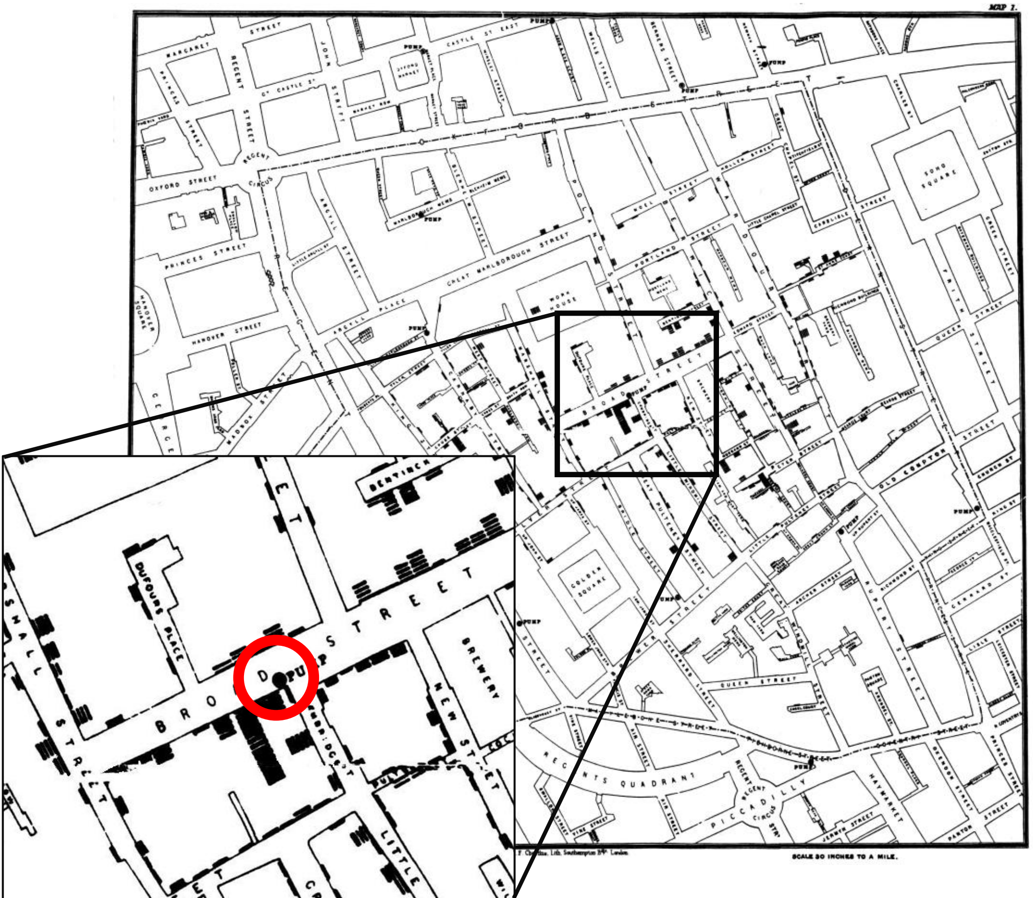
Potential exam questions for today’s lecture

- Explain what is map overlay. What different kinds of overlays exist for *vector* data?
- Explain the difference between a vector map overlay and spatial join
- What is spatial interpolation? Explain the use of TIN model in spatial interpolation. What kind of limitations relate to this method?
- Explain the method of Kernel density estimation. Kernel density estimation and spatial interpolation both create a surface from a point set. However, they are fundamentally different approaches. What is this difference? Give an examples, of what kind of datasets do they fit.
- Explain how to form a Voronoi diagram for a set of points. Draw a set of seven points and the Voronoi diagram for the set.

Mahdollisia tenttikysymyksiä tämän päivän luennosta

- Selitä mikä on map overlay. Millaisia erilaisia overlay-menetelmiä on vektorimuotoiselle datalle?
- Miten vektoridatan map overlay ja spatiaalinen liittäminen
- Mitä on spatiaalinen interpolointi? Selitä miten spatiaalinen interpolointi tehdään TIN-mallissa. Mitä rajoituksia menetelmällä on?
- Selitä Kernel-tiheysestimoinnin menetelmä. Kernel-tiheysestimointi ja spatiaalinen interpolointi molemmat tuottavat pinnan pistejoukon perusteella. Ne ovat kuitenkin oleellisesti erilaiset lähestymistavat. Mikä on tämä ero? Anna esimerkit, minkälaiselle datalle nämä lähestymistavat sopivat.
- Selitä kuinka Voronoi-diagrammi muodostetaan pistejoukolle. Piirrä seitsemän pisteen joukko ja sille Voronoi-diagrammi.

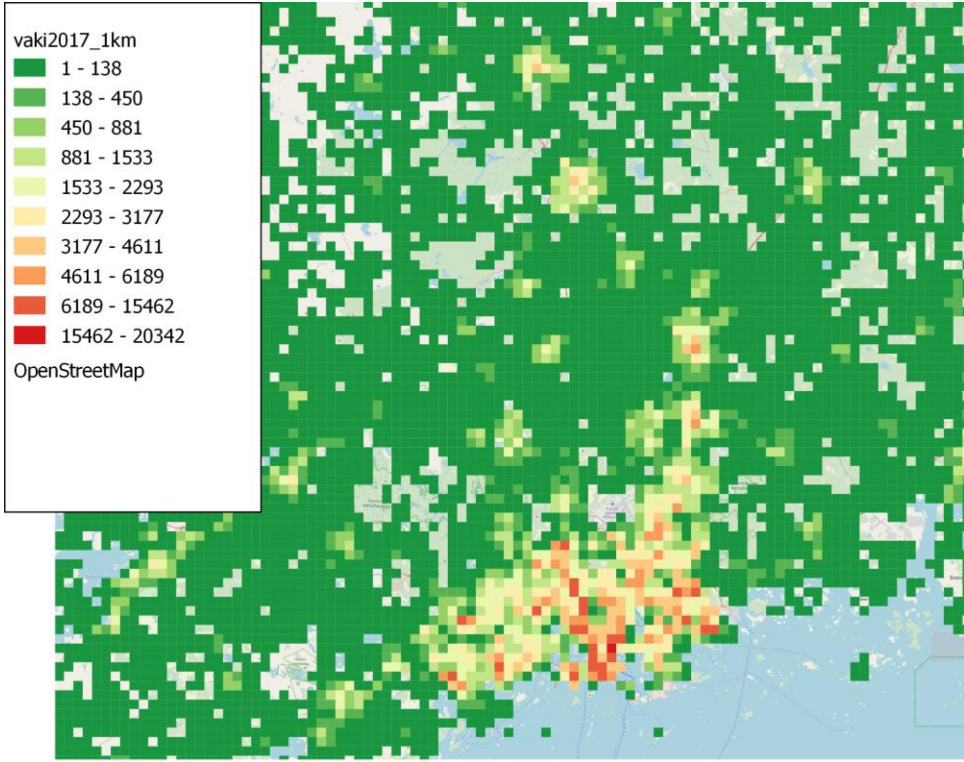
Analysis based on location



- **The idea of spatial analysis is to use location as central element of the analysis**
- **Shown here is what is commonly considered to be the first modern spatial analysis by Dr. John Snow about the 1854 Cholera epidemic in Soho, London**
  - The cause of the epidemic was a contaminated public water pump

Analysis based on location and attributes

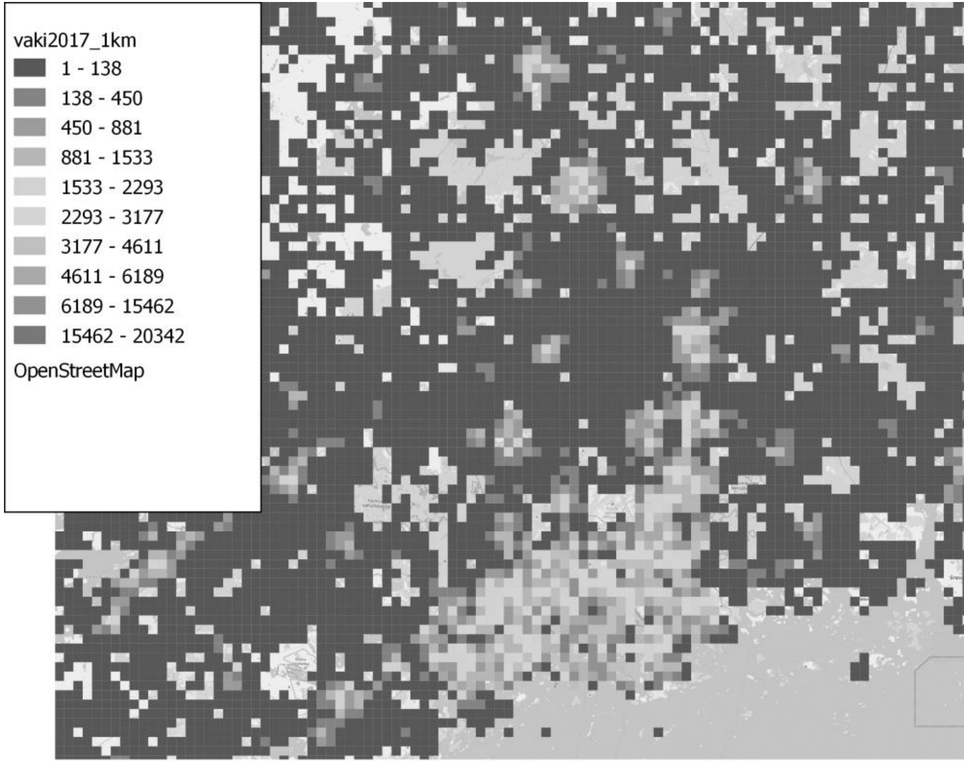
- **Spatial analysis combines location with attribute data in order to reveal patterns**
- **Example: vast majority of Finland is sparsely populated (less than 138 people/km^2)**
  - In this data set there are 35 762 km^2 where population density is 1-9 and 5 128 km^2 where population density is over 138
  - 99 528 km^2 are inhabited
  - Total Finnish land area is 308 891 km^2



Aalto University
School of Engineering

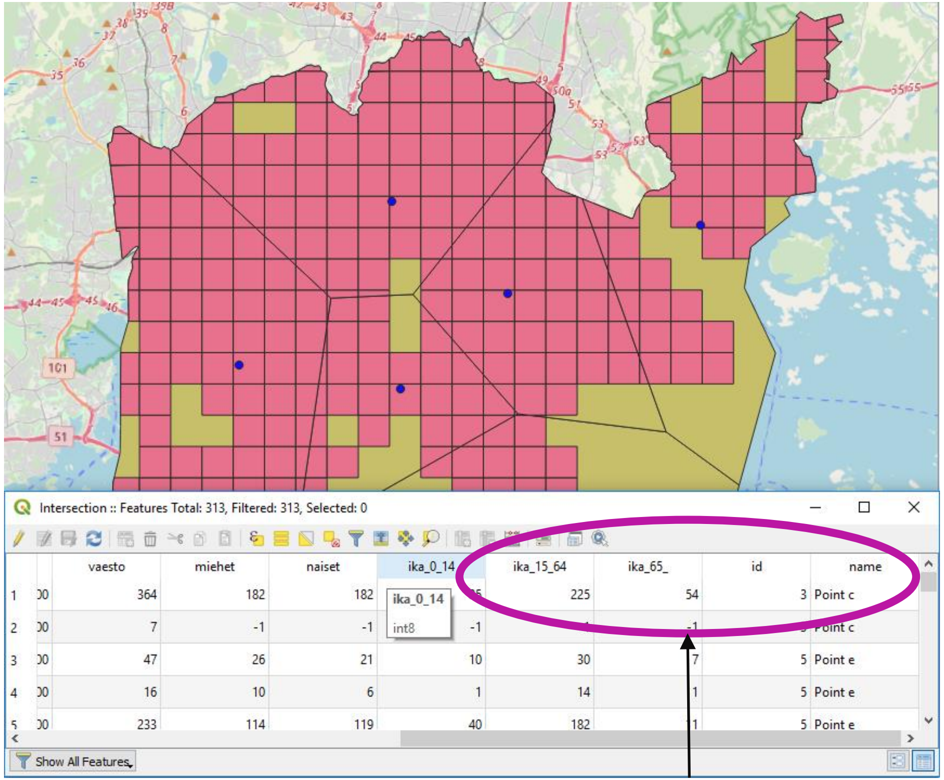
Analysis based on location and attributes

- **Spatial analysis combines location with attribute data in order to reveal patterns**
- **Example: vast majority of Finland is sparsely populated (less than 138 people/km^2)**
  - In this data set there are 35 762 km^2 where population density is 1-9 and 5 128 km^2 where population density is over 138
  - 99 528 km^2 are inhabited
  - Total Finnish land area is 308 891 km^2



Analysis based on several data layers

- **Spatial analysis is often based on several data layers**
- **Data layers can be combined in various ways, such as**
  - Map overlay
    - Create new layer based on input layers
  - Spatial or Table (attribute) join
    - Combine attribute data in existing layers
    - May also create new layers

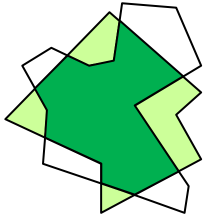


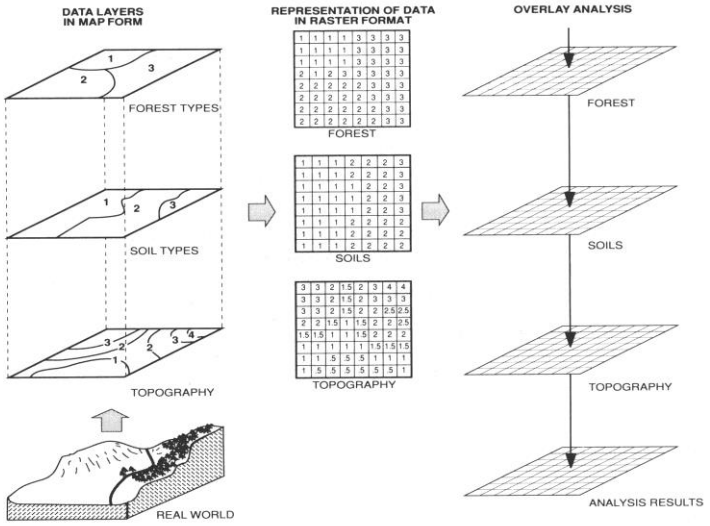
Data from different input layers

Map overlay and data layer joins

Map Overlay

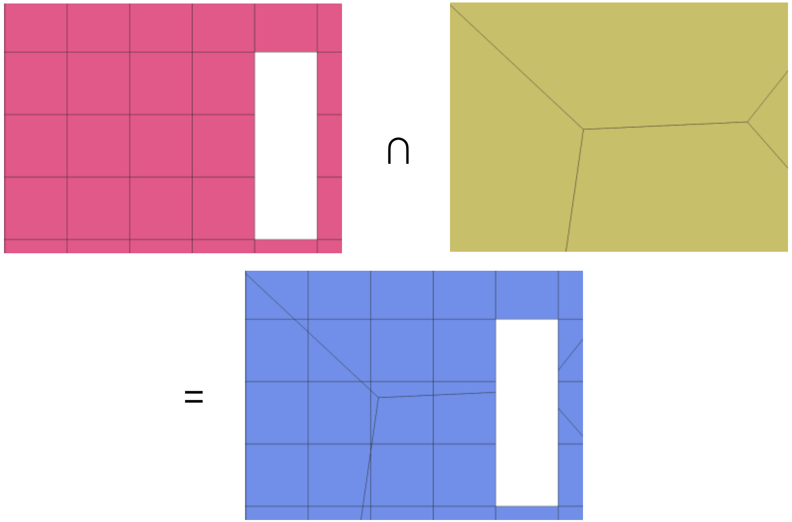
- One of the most basic spatial analysis methods is the map overlay
- Two or more data layers are put “on top of each other” and the result is a combination of data in the input layers
- Can be used for both raster and vector data





Vector Map Overlay

- For Vector data map overlay creates new polygons
- Output layer attributes typically are a combination of input data layer attributes
- There are several types of overlays that can be done for vector data



	ogc_fid	kunta	grd_id	id_nro	xkoord	ykoord	vaesto	miehet	naiset	ika_0_14	ika_15_64	ika_65_
19	19	091	1kmN6670E0393	47584	393000	6670000	308	158	150	73	192	43
20	20	091	1kmN6670E0394	47585	394000	6670000	53	25	28	12	36	5
21	21	049	1kmN6671E0378	48244	378000	6671000	369	189	180	82	214	73

+

	id	name
4	4	Point d
5	6	Point f
6	1	Point a

=

	ogc_fid	kunta	grd_id	id_nro	xkoord	ykoord	vaesto	miehet	naiset	ika_0_14	ika_15_64	ika_65_	id	name
4	266	091	1kmN6682E0401	55692	401000	6682000	16	10	6	1	14	1	5	Point e
5	265	091	1kmN6682E0400	55691	400000	6682000	233	114	119	40	182	11	5	Point e
6	264	091	1kmN6682E0399	55690	399000	6682000	304	146	158	69	229	6	5	Point e

Different Overlay operations for Vector Data

- **Most common vector overlay operations are**
  - Intersect: retains overlapping parts
  - Union: retains everything
  - Subtract[1]: retain non-overlapping parts from one layer
- **Existing polygons are modified and new created as required**
  - May also cause changes in attribute data

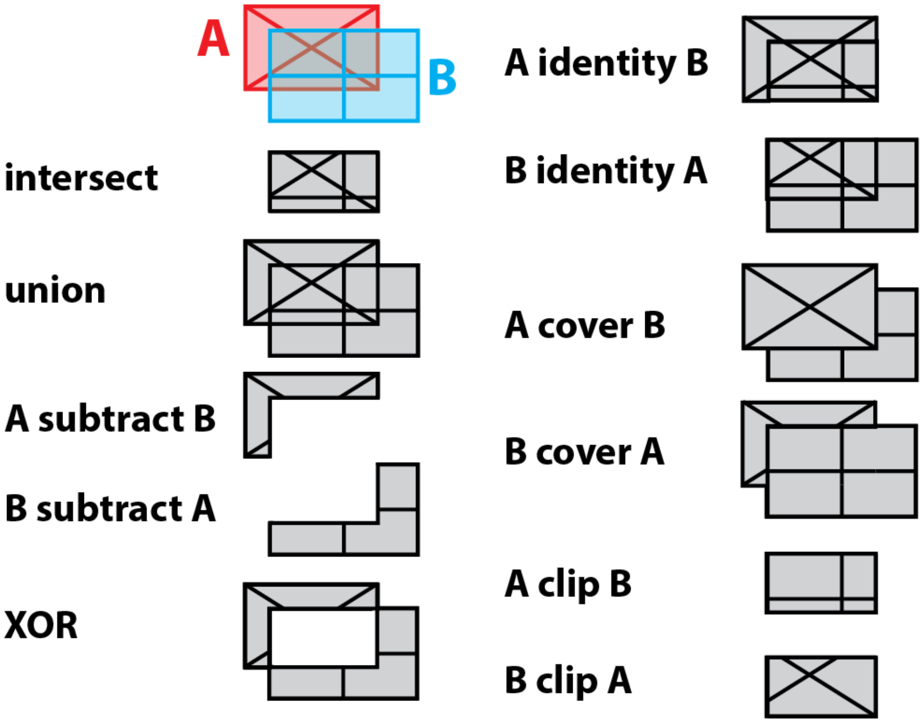


Image source: http://wiki.gis.com/wiki/index.php/Overlay

Aalto University
School of Engineering

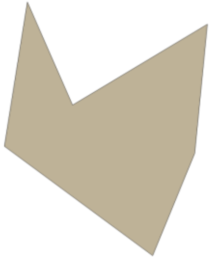
[1]Also called difference, erase

https://presemo.aalto.fi/enyc2005/

Classroom exercise: overlay examples

Input data layers

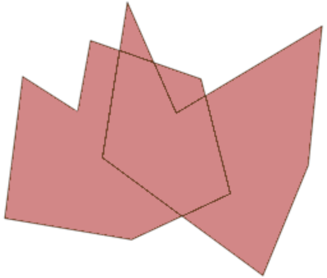
&



⟹

1.

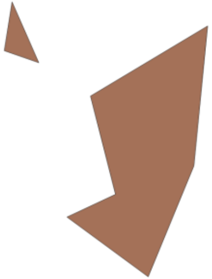
2.



A

C

B



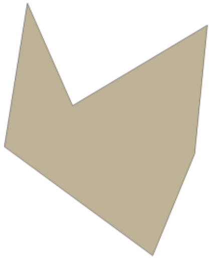
D

Name the overlay operation: intersection, union, subtraction, exclusive or, identity, cover, or clip?

Classroom exercise: overlay examples

Input data layers

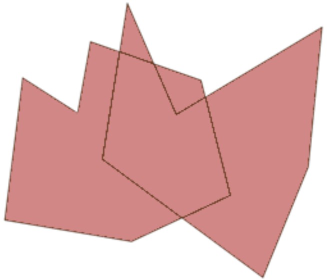
&



⟹

1.

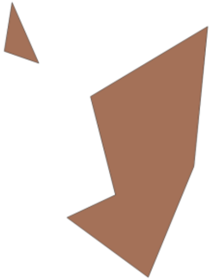
2.



A: Union

C: Intersection or clip[1]

B: 1 subtract 2



D: 2 subtract 1

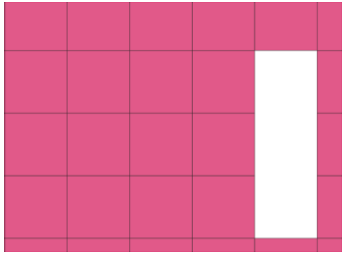
Name the overlay operation: intersection, union, subtraction, exclusive or, identity, cover, or clip?

[1]with this data, the results of the two operations look identical

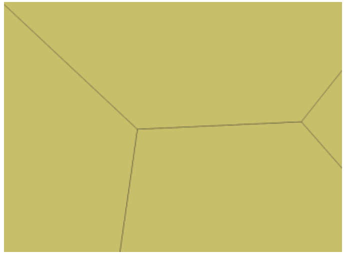
Aalto University
School of Engineering

Spatial and Table Joins

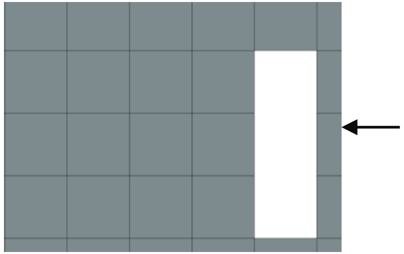
- **Join combines the attribute data from two layers**
  - In spatial join the new attribute data is selected by location
  - In table join, the attributes are joined according to a key value
    - The joined table does not need to contain spatial data



join



=



Polygons do not change

	ogc_fid	kunta	grd_id	id_nro	xkoord	ykoord	vaesto	miehet	naiset	ika_0_14	ika_15_64	ika_65_
19	19	091	1kmN6670E0393	47584	393000	6670000	308	158	150	73	192	43
20	20	091	1kmN6670E0394	47585	394000	6670000	53	25	28	12	36	5
21	21	049	1kmN6671E0378	48244	378000	6671000	369	189	180	82	214	73

+

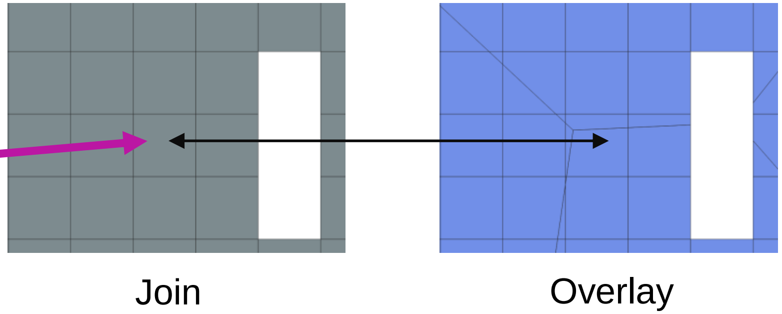
	id	name
4	4	Point d
5	6	Point f
6	1	Point a

=

	ogc_fid	kunta	grd_id	id_nro	xkoord	ykoord	vaesto	miehet	naiset	ika_0_14	ika_15_64	ika_65_	id	name
4	266	091	1kmN6682E0401	55692	401000	6682000	16	10	6	1	14	1	5	Point e
5	265	091	1kmN6682E0400	55691	400000	6682000	233	114	119	40	182	11	5	Point e
6	264	091	1kmN6682E0399	55690	399000	6682000	304	146	158	69	229	6	5	Point e

Joins can be one-to-one or one-to-many; contents of the result table depend on this

Spatial and Table Joins



- **In one-to-one join this input polygon is in the output once**
  - Joined with one of the three polygons it overlaps with
- **In one-to-many join the input polygon is in the output three times**
  - Once for each polygon it overlaps with

One-to-one join

ogc_fid	kunta	grd_id	id_nro	xkoord	ykoord	vaesto	miehet	naiset	ika_0_14	ika_15_64	ika_65_	id	name
111	091	1kmN6676E0385	51626	385000	6676000	764	352	412	108	435	221	2	Point b
[illegible]	091	1kmN6676E0386	51627	386000	6676000	4325	1892	2433	637	2989	699	[illegible]	[illegible]
113	09	1kmN6676E0387	51628	387000	6676000	5879	2637	3242	803	4379	697	6	Point f
[illegible]	091	1kmN6676E0388	51629	388000	6676000	2448	1129	1319	551	1537	360	[illegible]	[illegible]

One-to-many join

ogc_fid	kunta	grd_id	id_nro	xkoord	ykoord	vaesto	miehet	naiset	ika_0_14	ika_15_64	ika_65_	id	name
112	091	1kmN6676E0386	51627	386000	6676000	4325	1892	2433	637	2989	699	3	Point c
113	[illegible]	1kmN6676E0387	51628	387000	6676000	5879	2637	3242	803	4379	697	6	Point f
113	09	1kmN6676E0387	51628	387000	6676000	5879	2637	3242	803	4379	697	2	Point b
113	09	1kmN6676E0387	51628	387000	6676000	5879	2637	3242	803	4379	697	3	Point c

Spatial and Table joins

- **Table join combines the contents of two layers' attribute tables based on shared attribute IDs**
- **In the example, the joined table (on right) contains land use types, which are joined to the objects in the left table**

ID	Land use code
1	1
2	2
3	1
4	3
5	2

Join

Land use code	Land use type
1	Built environment
2	Agricultural land
3	Water

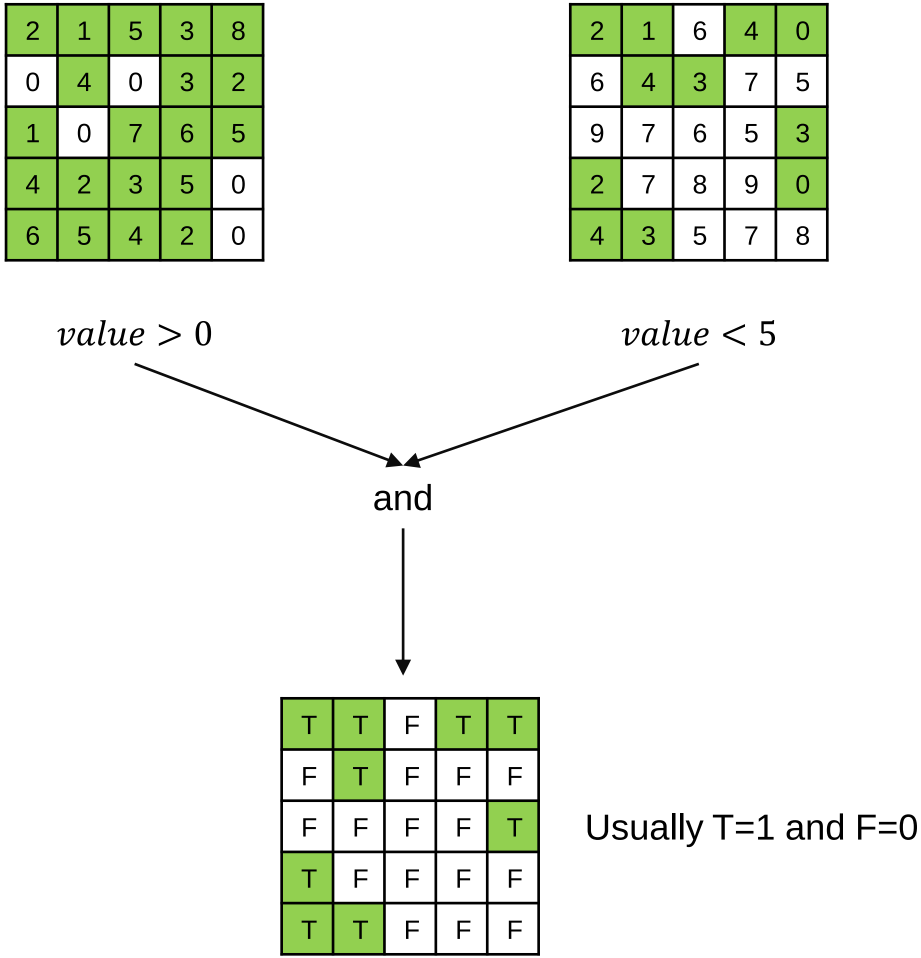
=

ID	Land use code	Join_Land use code	Land use type
1	1	1	Built environment
2	2	2	Agricultural Land
3	1	1	Built environment
4	3	3	Water
5	2	2	Agricultural land

The attribute named "Join_Land use code" in the result table is used to distinguish between the two attributes with identical names in the input tables

Map overlay in raster data

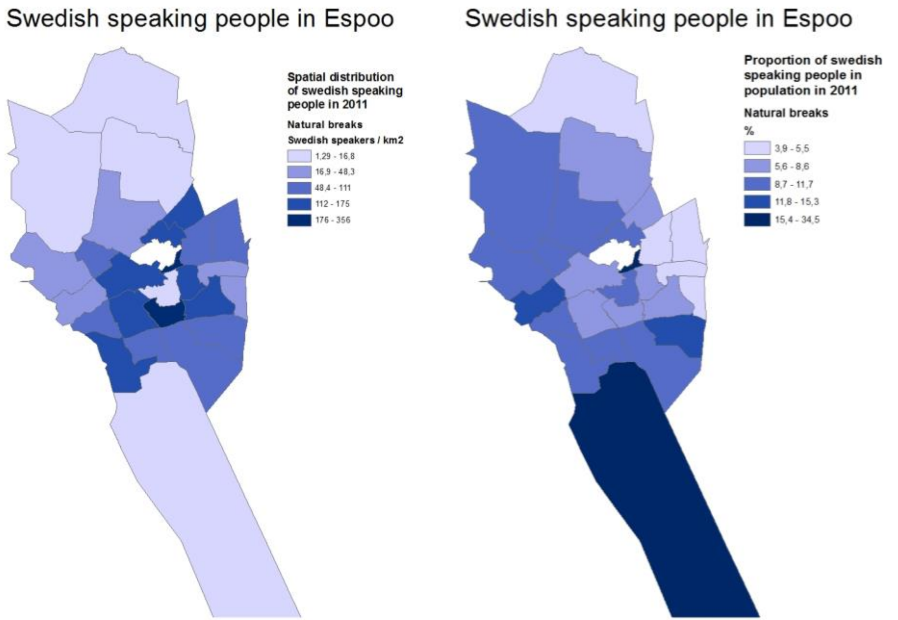
- For raster data map overlay combines cell values at the same location in some manner
- Simplest overlay combines binary values (true/false)
- More details later on the course



Data normalization

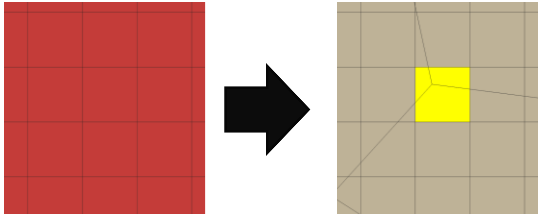
Location, area, and attribute values

- **Attribute values in spatial data are dependent on the location**
- **Thus, both geographic data and attribute data need to be considered when handling spatial data sets**
    - Data comparison to geographic or attribute data?



Data normalization in analysis processes

- **Spatial analysis processes can modify the geometry of elements**
  - These modifications may not be automatically reflected in attribute values
  - Attributes need to be modified to reflect the new situation

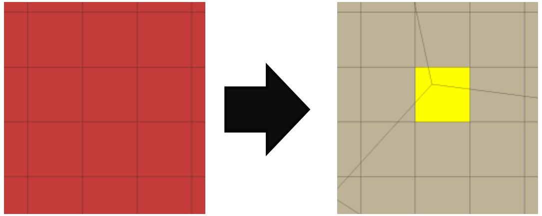


ogc_fid	kunta	grd_id	id_nro	xkoord	ykoord	vaesto
90	091	1kmN6675E0385	50951	385000	6675000	3997

ogc_fid	address	kunta	grd_id	id_nro	xkoord	ykoord	vaesto
53	Keskuspelastus...	091	1kmN6675E0385	50951	385000	6675000	3997
69	Haagan pelastu...	091	1kmN6675E0385	50951	385000	6675000	3997
107	KÃ¤pylÃ¤n pel...	091	1kmN6675E0385	50951	385000	6675000	3997

Data normalization in analysis processes

- **Proper normalization depends on the situation**
- **In this case: area divided into smaller pieces ⇒ population needs to be adjusted**
  - Adjustment according to new area is appropriate if no other information is available

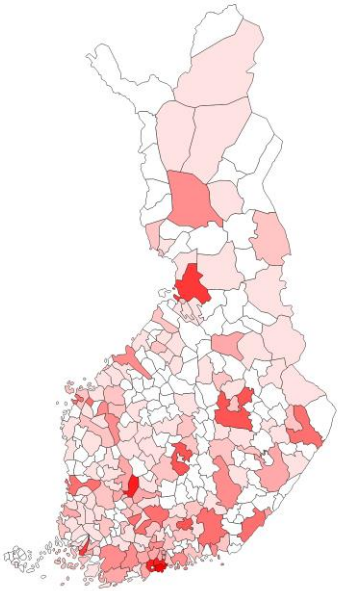


ogc_fid	kunta	grd_id	id_nro	vaesto
90	091	1kmN6675E0385	50951	3997

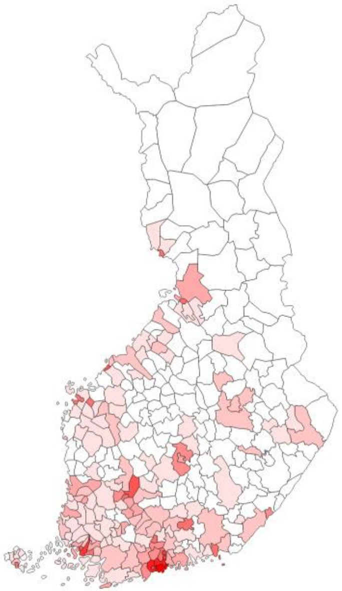
ogc_fid	address	kunta	grd_id	id_nro	vaesto	areafract	populatio
107	KÃ¤pylÃ¤n pel...	091	1kmN6675E0385	50951	3997	0,258	1031
69	Haagan pelastu...	091	1kmN6675E0385	50951	3997	0,136	544
53	Keskuspelastus...	091	1kmN6675E0385	50951	3997	0,605	2418

Normalization in data representation

- **Representation or comparison of “raw” values is often not very useful**
  - Data needs to be normalized for it to have a reasonable meaning
  - Normalization can depend on the spatial characteristics
  - Proper normalization is important in order to be able to properly interpret the results
  - Visualization also plays a major role



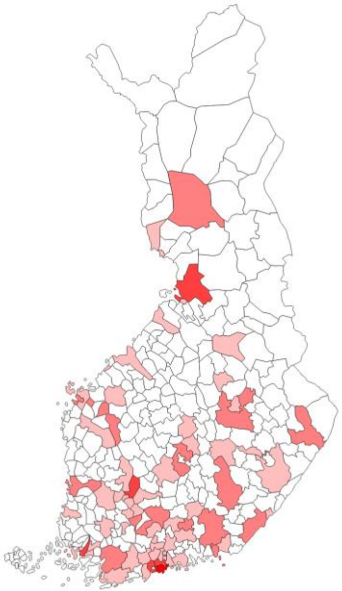
Population of Finnish municipalities (10 classes)



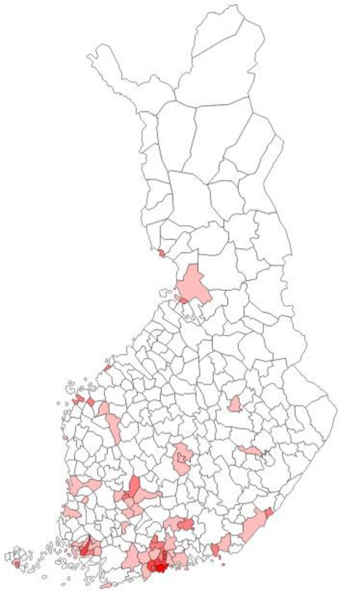
Population density of Finnish municipalities (10 classes)

Normalization in data representation

- **Representation or comparison of “raw” values is often not very useful**
  - Data needs to be normalized for it to have a reasonable meaning
  - Normalization can depend on the spatial characteristics
  - Proper normalization is important in order to be able to properly interpret the results
  - Visualization also plays a major role



Population of Finnish municipalities (5 classes)



Population density of Finnish municipalities (5 classes) https://presemo.aalto.fi/enyc2005/

Classroom exercise: rescue service preparedness

- **Consider how rescue services for Helsinki prepare for incidents**
  - What information is needed for planning rescue services?
  - How location affects the information?

Aalto University
School of Engineering

Images: wikimedia

Classroom exercise: rescue service preparedness

- Incident **density**
- Population **density**
- Road network and **time-to-location**
- Effect of day/night-time
- Distribution of resources
- Locations for operations
- Etc.

Images: wikimedia

From points to surfaces: density surfaces and interpolated surfaces

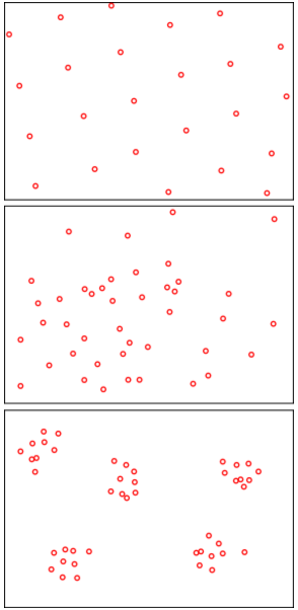
Points to surfaces

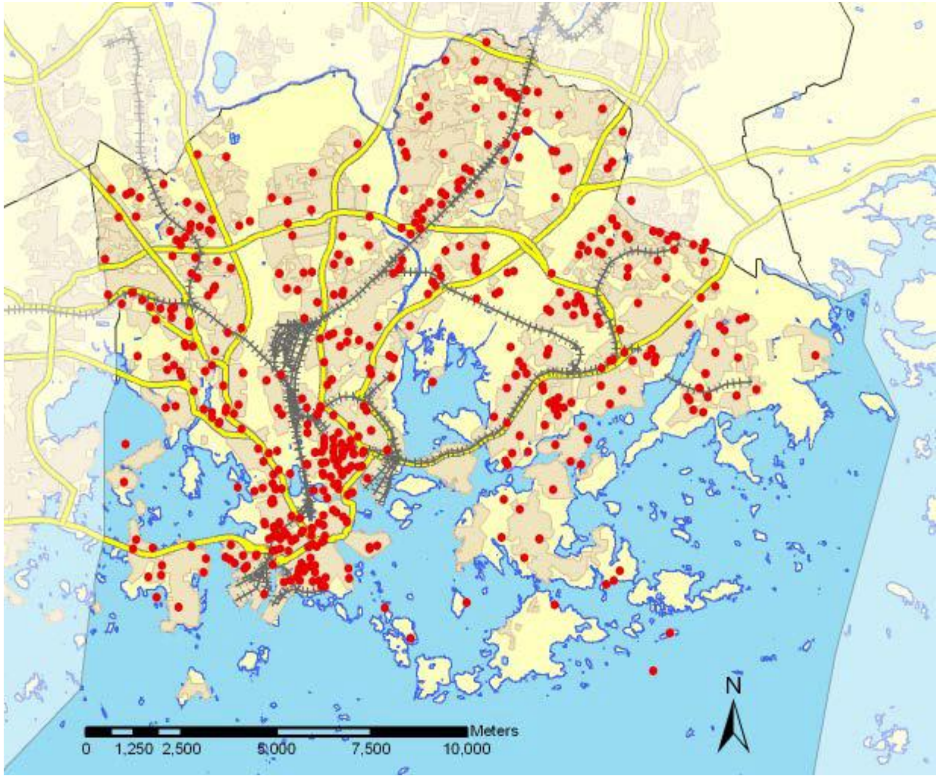
- **Data transformations are often required in spatial analysis**
- **Generalization**
  - E.g. points to density surfaces
- **Interpolation**
  - Unknown values derived from known values
- **More complex analysis**
  - Service areas, effect spreading from a point source, etc.

Points to density surface: rescue service case revisited

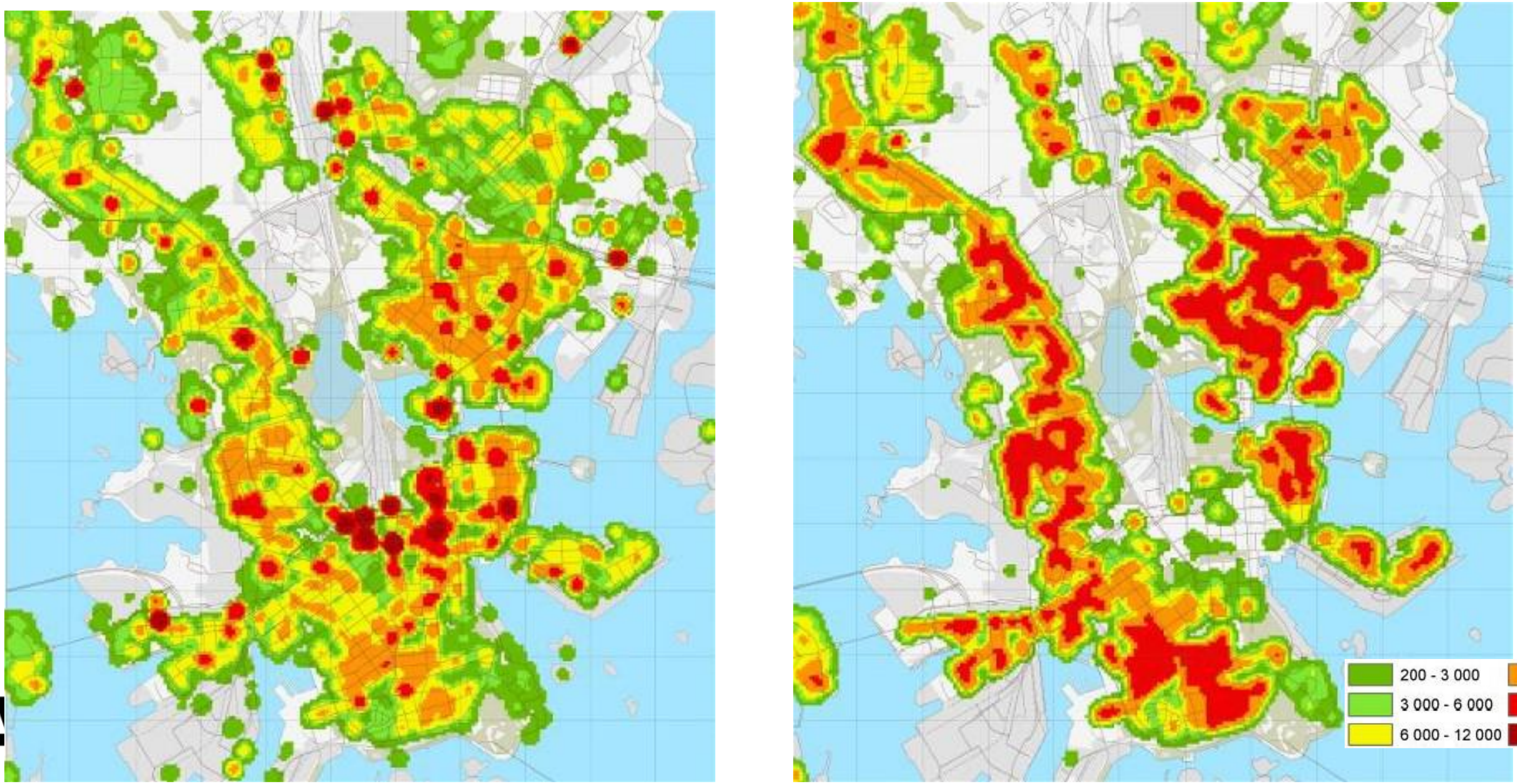
Incidents of domestic fires in Helsinki

The distribution of the events is neither regular nor random



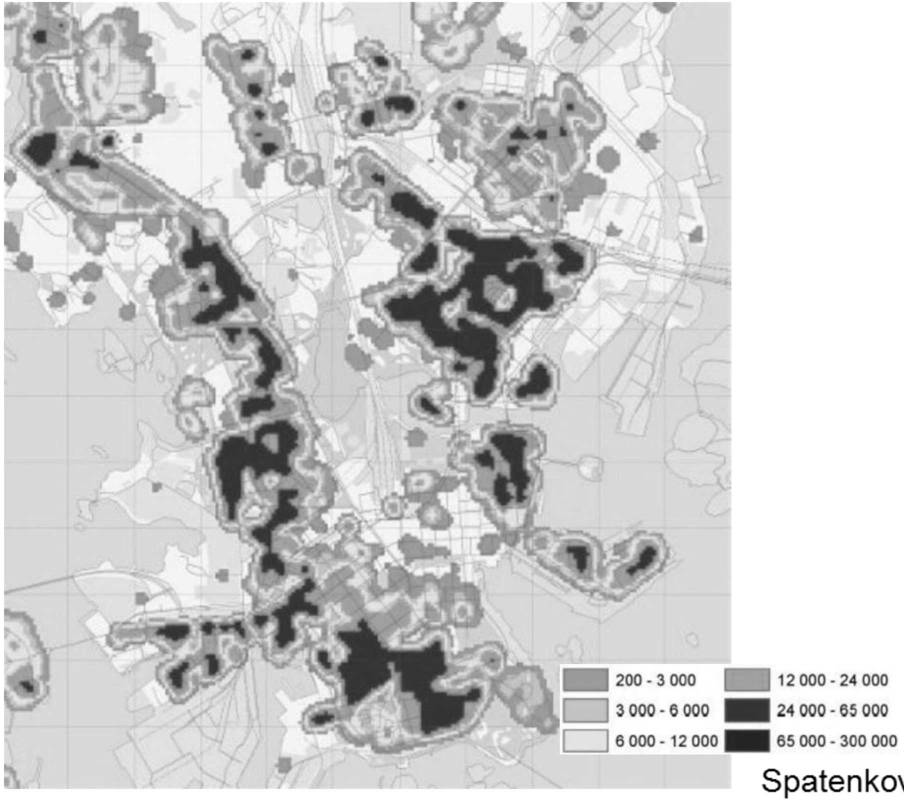


Points to density surface: rescue service case revisited



Spatenkova 2009

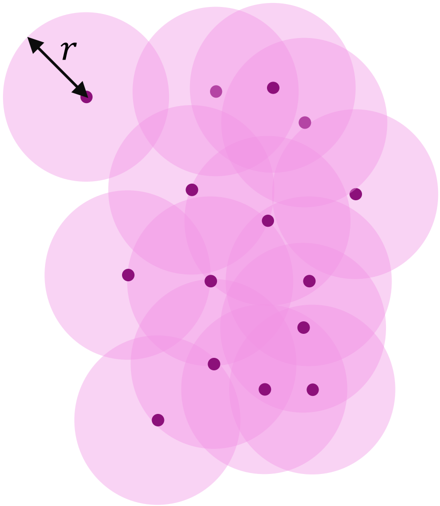
Points to density surface: rescue service case revisited



Spatenkova 2009

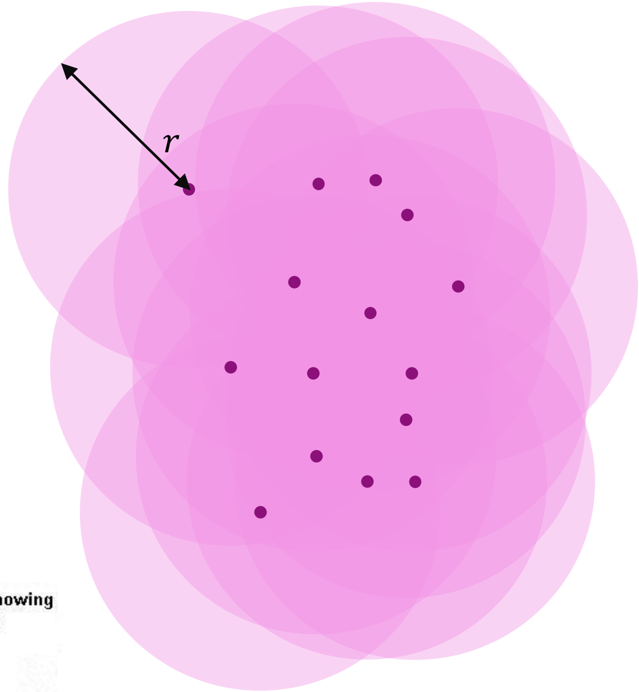
Kernel function

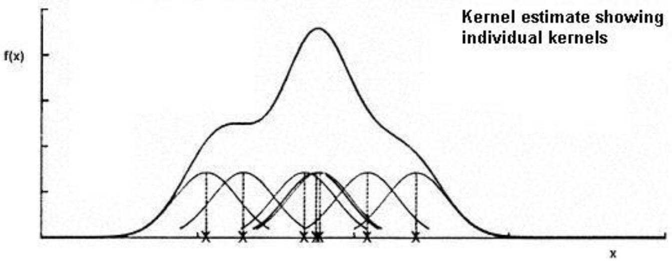
- **Kernel function transforms point data to a density surface**
  - Visualization
  - Comparison of point and surface data sets
  - Gives a density at any location
- **Simple kernel is a circular buffer with constant value on radius r around each point p**



Kernel function

- **The bandwidth (radius) of the kernel affects the resulting surface**
  - Large bandwidth ⟹less variation
  - Small bandwidth ⟹ more variation
- **More sophisticated kernel function weights distance from the point**



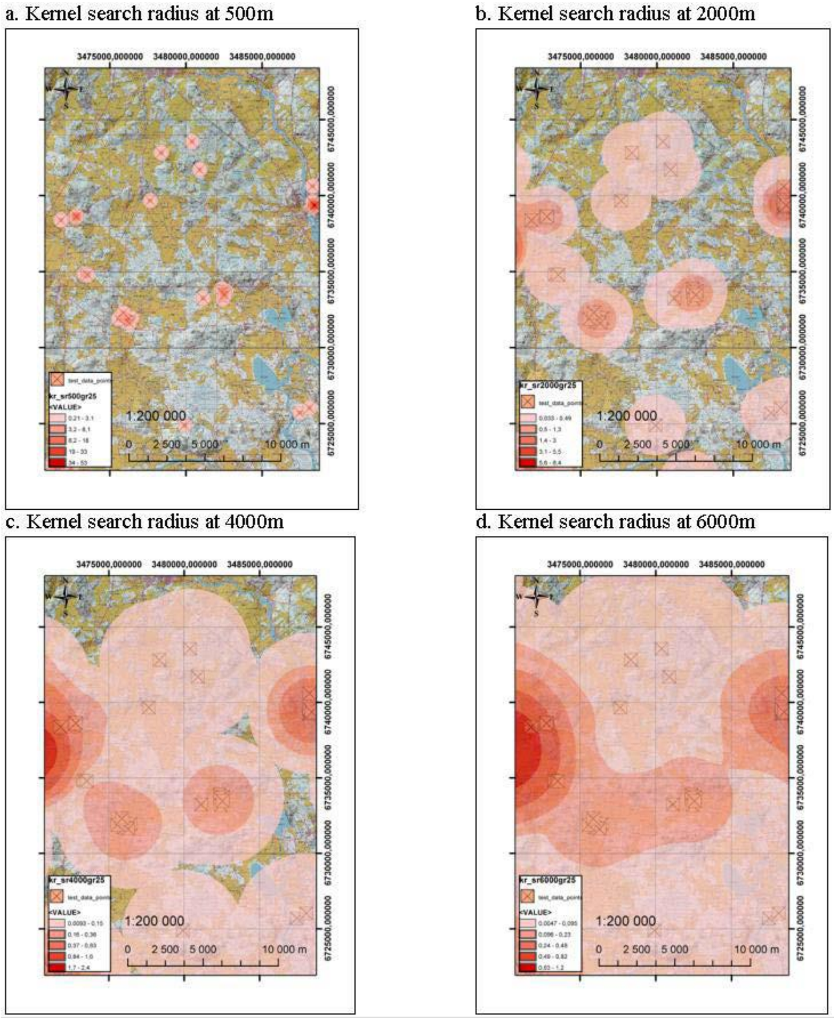


Krisp 2006

Aalto University
School of Engineering

Kernel function

- Kernel method results depend heavily on the interpretation
- Requires expertise from the user: what is a proper bandwidth?
- Visualization, again, can affect the interpretation a lot



Point density and spatial interpolation

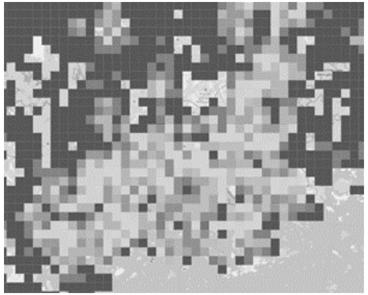
- **Density surface estimates how many events (data points) there are per area**
  - Makes sense when data describes discrete elements
  - Population density, incident density, lightning strike density, etc.
  - Object model data
- **What about a situation where the points represent value of a field phenomenon at specific locations?**
  - Temperature density, or elevation density does not make sense
- **Interpolation is a method for estimating value of a field phenomenon at locations where it is not known**

Spatial autocorrelation

- **Value of a field phenomenon can be interpolated because of spatial autocorrelation**
- **Values of a field phenomenon are likely to be similar, the closer to each other they are**
  - Usually

- **Tobler's first law of geography: everything is related to everything else, but near things are more related than distant things**

Population density gradually decreases farther away from Helsinki city centre



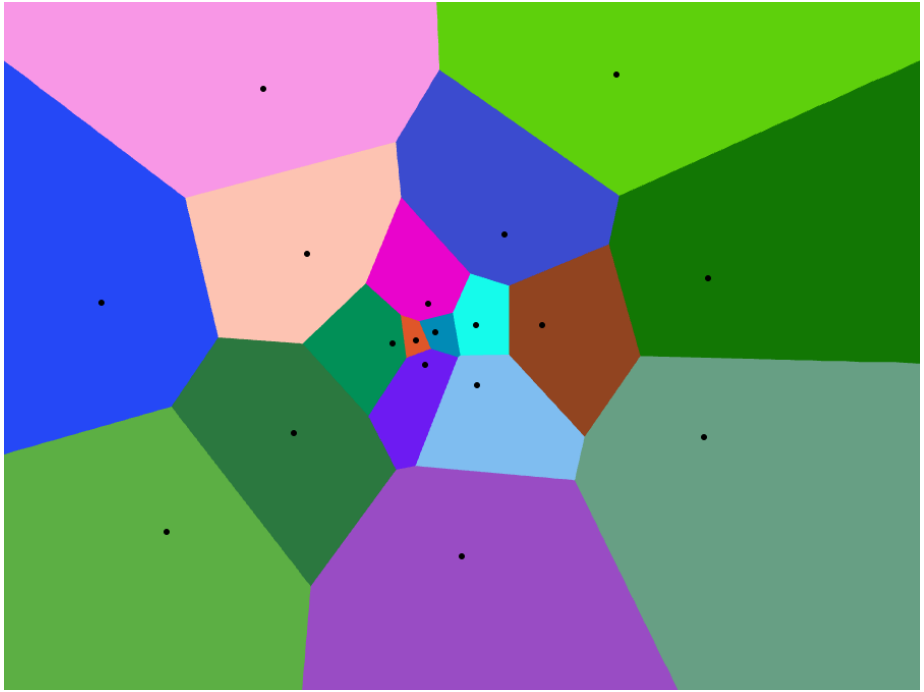
Except when you go south (where it quickly drops to 0)

Spatial autocorrelation and interpolation

- **Spatial interpolation is possible with positive spatial autocorrelation**
  - **Nearby** known values can be used to estimate unknown values
- **Plenty of different methods**
  - Local or global, deterministic or stochastic, smooth or abrupt…
- **Voronoi diagram: values nearest to the point to be interpolated**
- **TIN model: linear interpolation from triangle corners**
- **Local spatial average, inverse distance weighing, Kriging, etc.**

Voronoi diagrams

- **For a set of points S on a plane, the Voronoi diagram is subdivision of the plane into areas, where for each point of S, the face (cell) surrounding it contains the area that is closer to it than to any other point in the set**
- **Formally:**
  - $\{p_1 \dots p_n\}$ is a finite set of points S on a 2d plane P
  - $d(x, y) \geq 0$ is a distance function
  - Voronoi polygon $V(pi)$ for $p_i \in S$ is defined as the set of locations $p \in P$ where $d(p, pi) \leq d(p, p_j) \forall j, i \neq j$



Remember: colors are there only for visualization!

Generated using http://alexbeutel.com/webgl/voronoi.html

Simple Voronoi interpolation

- **Value within a Voronoi cell (proximity polygon) is the value of the point $p \in S$ used to define the cell**
- **Value changes abruptly when moving from one cell to another**
- **If data is nominal (categorical), this may be a good approach**

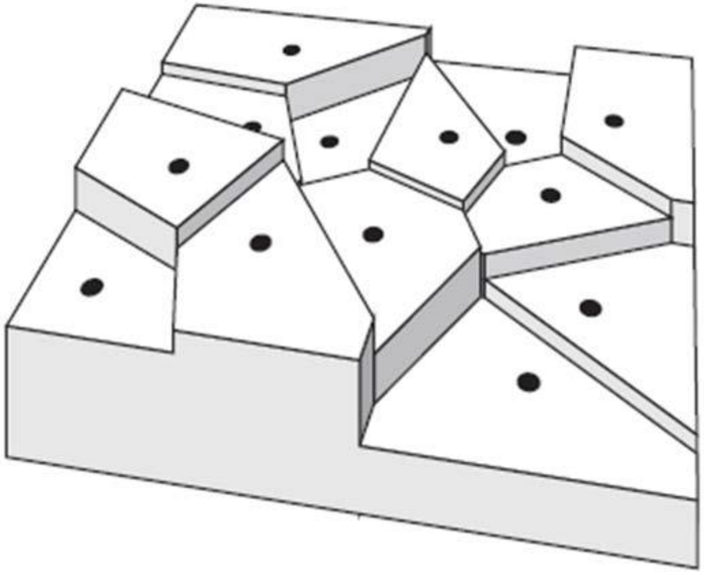
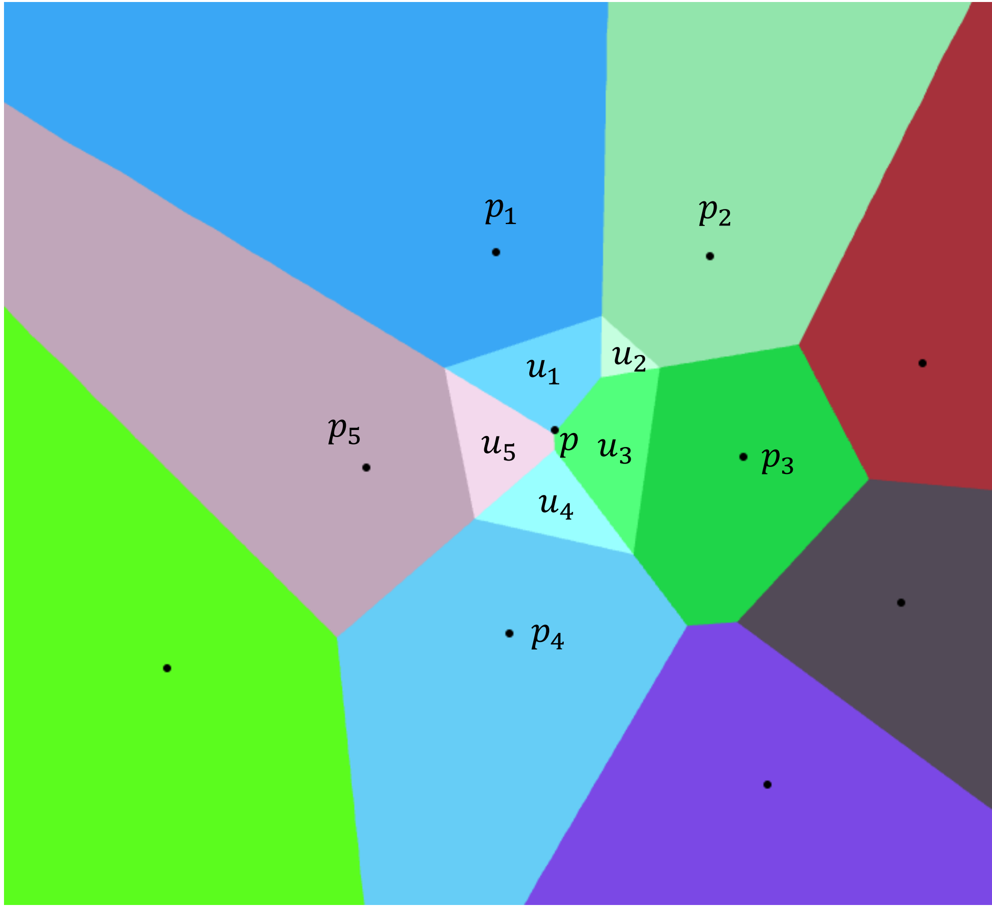


Image source: O'sullivan & Unwin (2010) Fig. 9.6 (p. 254)

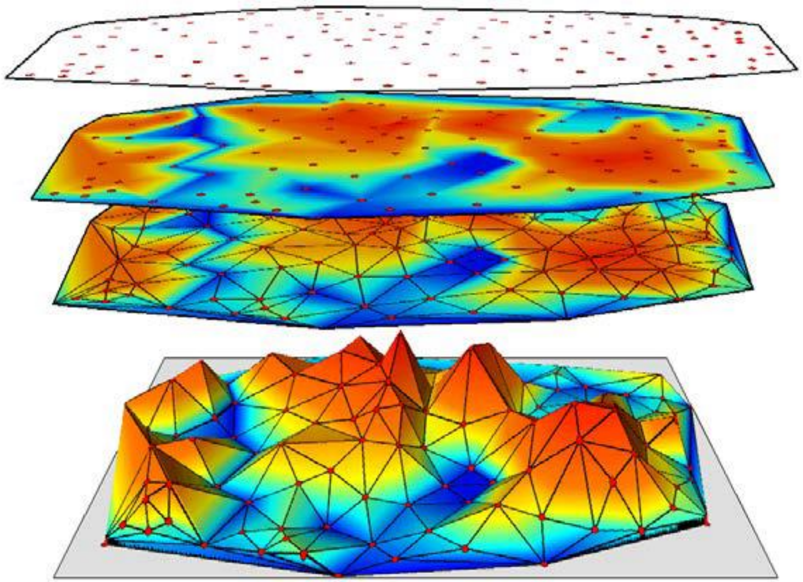
Sibson (natural neighbor) interpolation

- Each point in the point set $S = \{p_1 \ldots p_n\}$ measures a value at the given location
- We would like to know the value at a point $p \notin S$
- $f(p) = \frac{\sum_{i=1}^{k} u_i f(p_i)}{\sum_{i=1}^{k} u_i}$
- The value of each neighboring point p_i weighted by the proportion of the area of $V(p)$ taken from $V(p_i)$
- Not good for creating surfaces; good for single points



TIN as a surface model

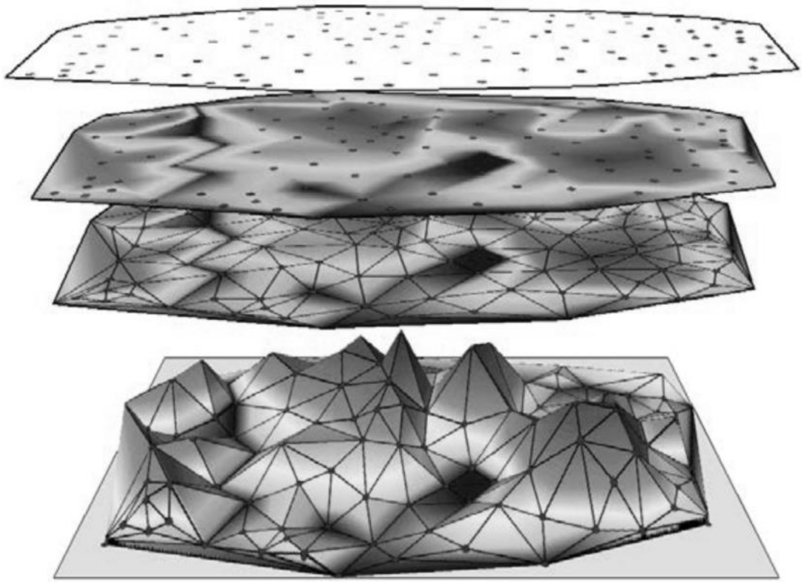
- **A TIN represents a surface calculated from point set S**
- **Each triangle is a plane, where value is based on values at triangle vertices (corners)**
  - Linear interpolation
- **Abrupt changes possible at triangle boundaries**
- **TIN is a surface**



www.geosolutions.com

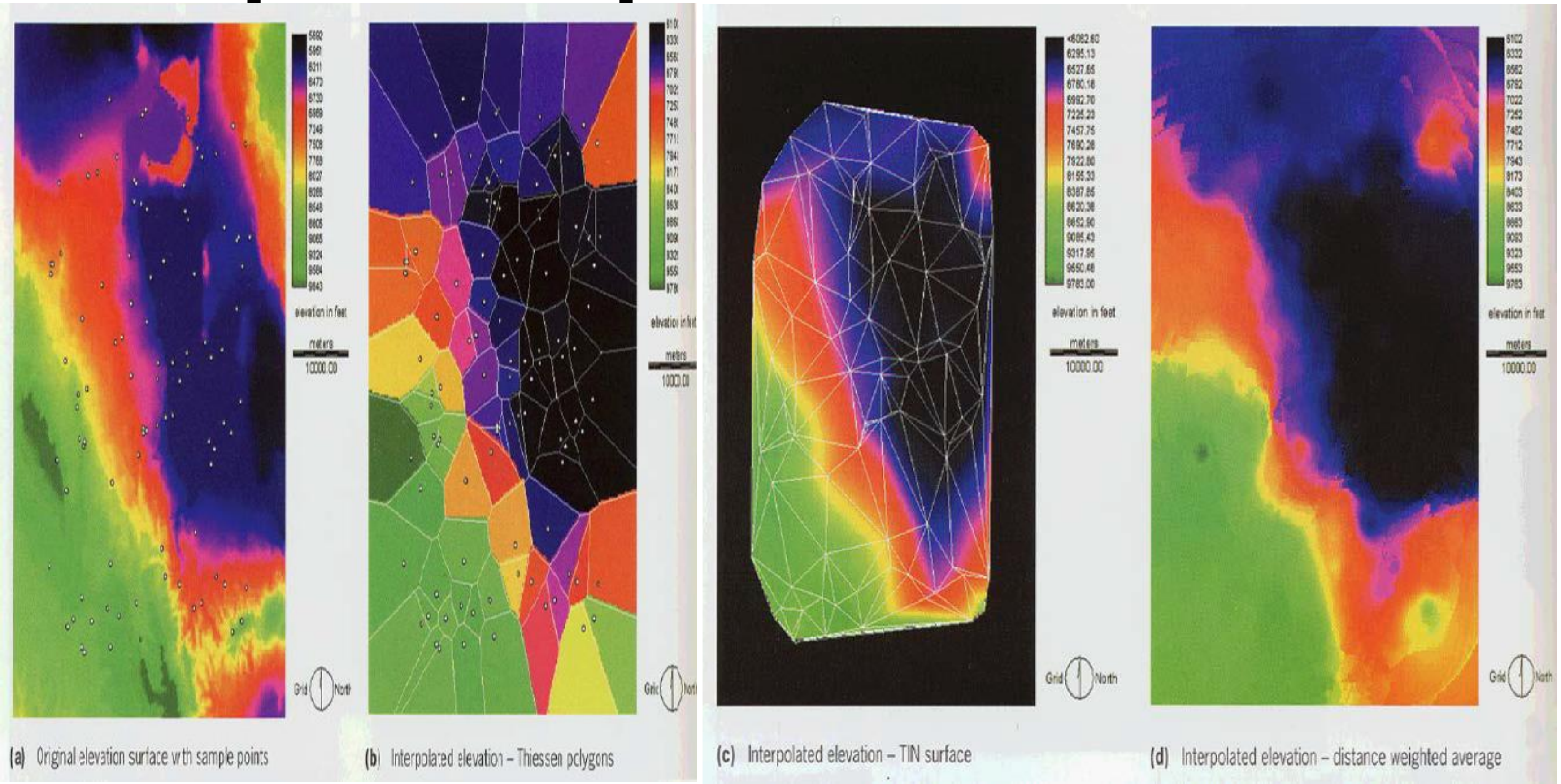
TIN as a surface model

- **A TIN represents a surface calculated from point set S**
- **Each triangle is a plane, where value is based on values at triangle vertices (corners)**
  - Linear interpolation
- **Abrupt changes possible at triangle boundaries**
- **TIN is a surface**



www.geosolutions.com

Example interpolation results

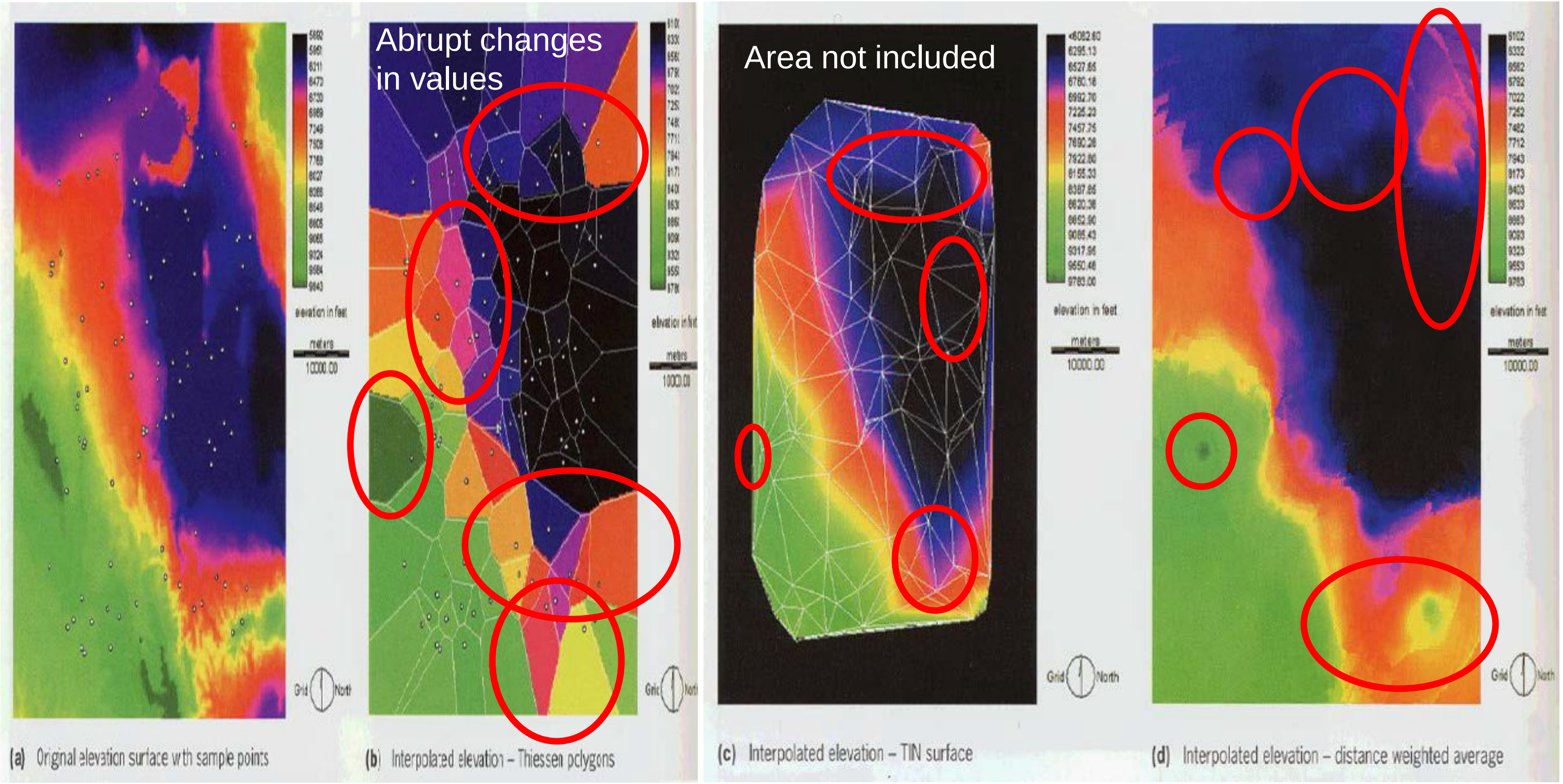


(a) Original elevation surface with sample points
(b) Interpolated elevation – Thiessen polygons
(c) Interpolated elevation – TIN surface
(d) Interpolated elevation – distance weighted average

Original surface | Voronoi diagram | TIN surface | IDW

Heywood et al 2003

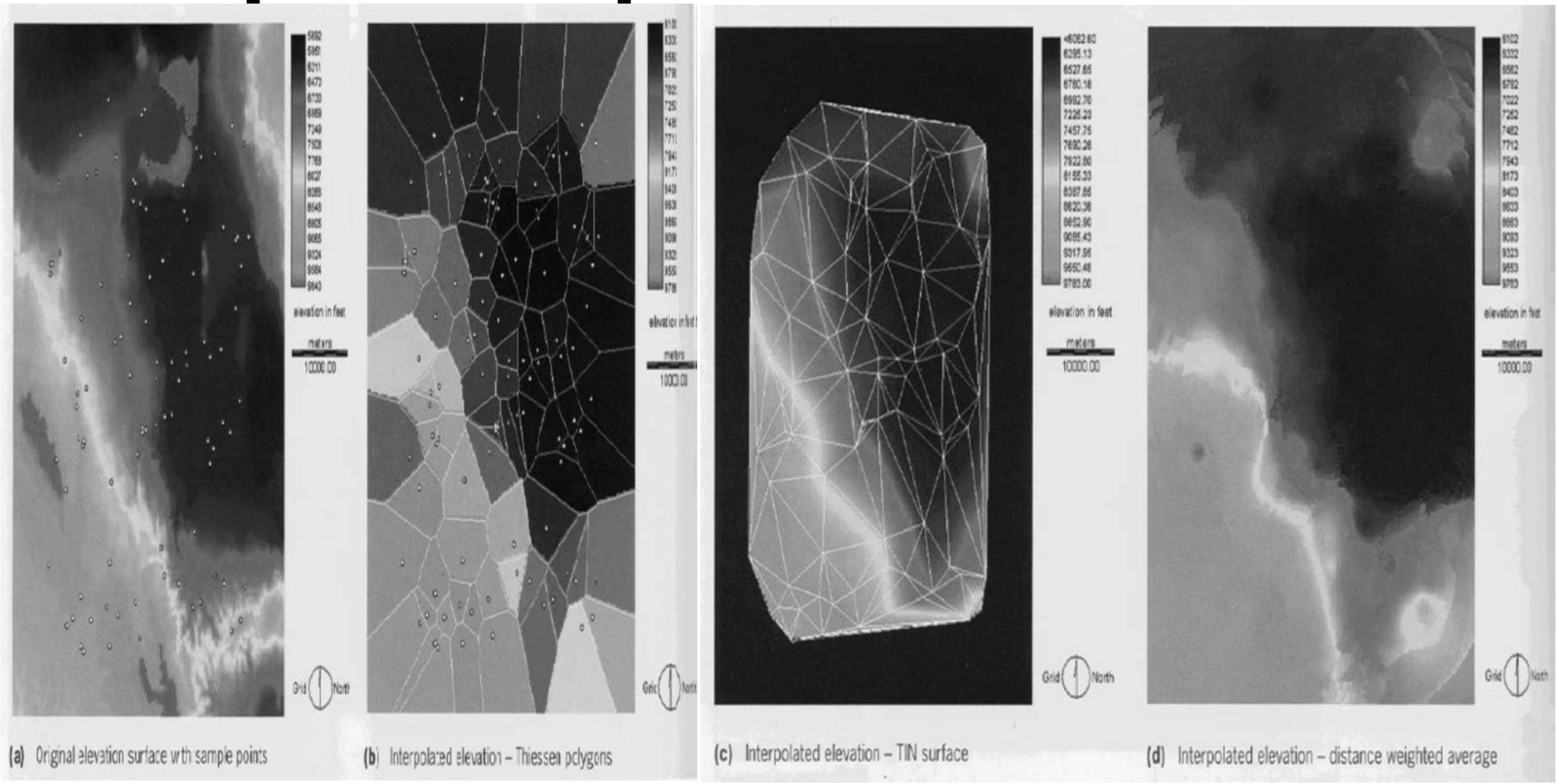
Brief assessment of the results



Original surface | Voronoi diagram | TIN surface | IDW

Heywood et al 2003

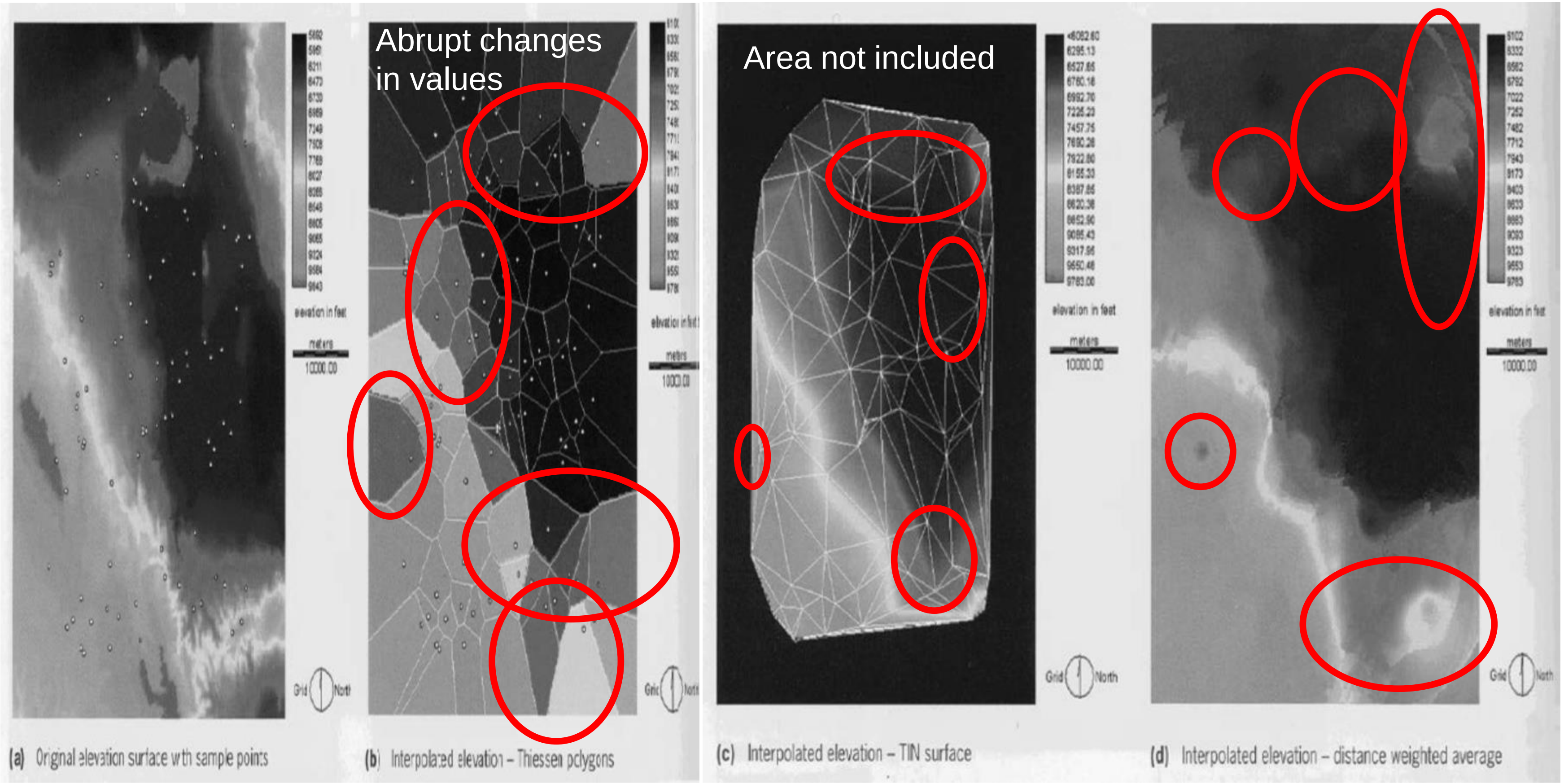
Example interpolation results



Original surface | Voronoi diagram | TIN surface | IDW

Heywood et al 2003

Brief assessment of the results

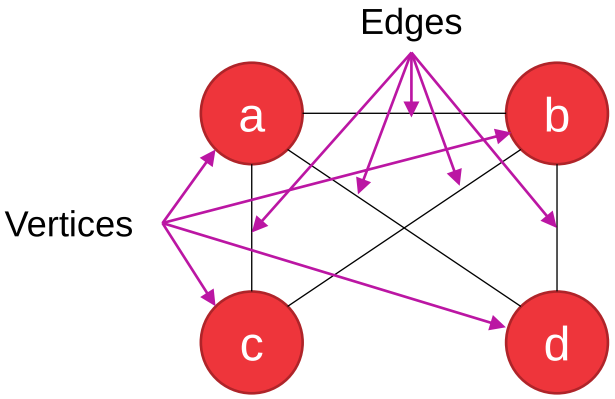


Original surface | Voronoi diagram | TIN surface | IDW

Heywood et al 2003

Network analysis

Network analysis



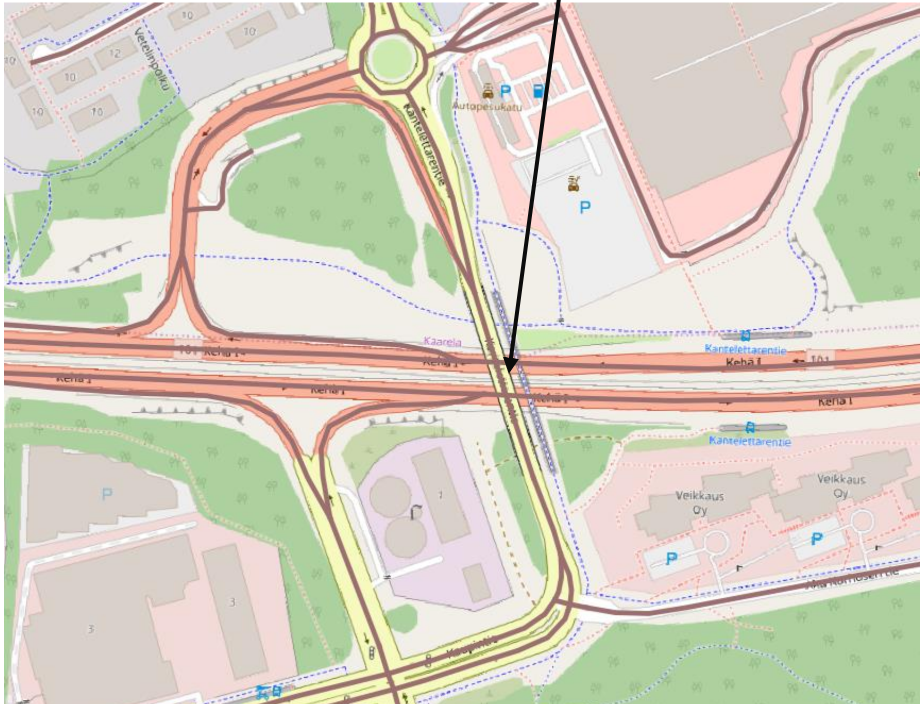
- **Graph (network) is a collection of vertices (nodes) connected by edges (links, arcs) mathematically $G = (V, E)$**
  - G is graph, V is a set of vertices $u \in V$, and E is a set of edges $(u, v) \in E$
- **Vertices can hold data values (attributes)**
- **Edges can also have attributes, such as**
  - Direction ((a, b) does not imply (b, a))
  - Weight (depicts, for example, distance or strength of connection between vertices)

- **In many spatial applications it's the edges that are important**
  - Road network, electricity network, water network, etc.

Network data example: road data

- **Roads are a common example of data that can be represented as a network**
- **Just having the physical shape of the network (polylines) is not sufficient**
- **A network requires explicit connections between road segments (= vertices)**

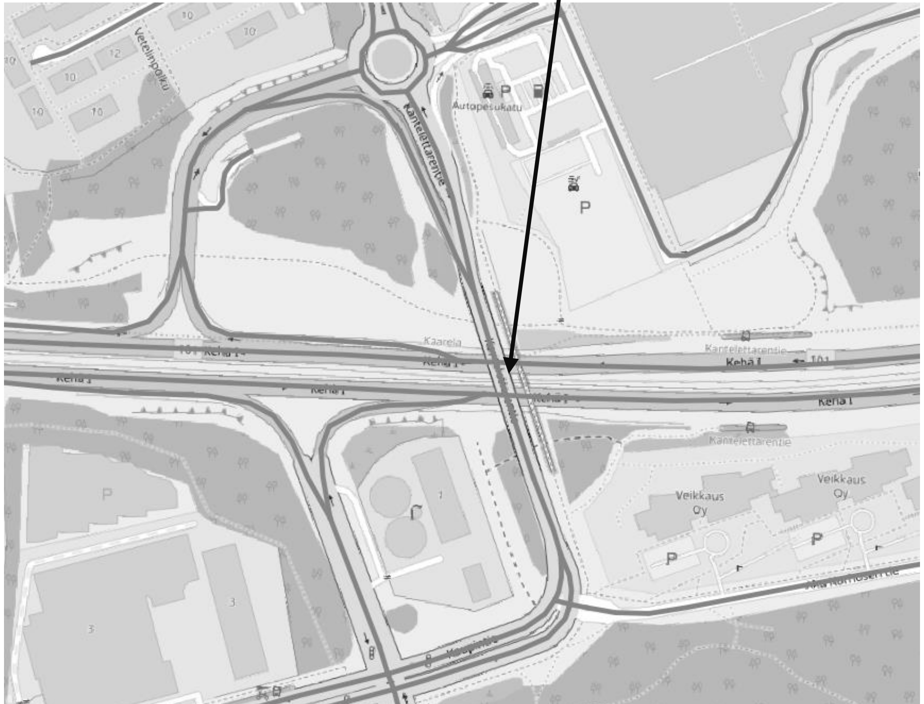
There is no connection here (overpass)



Network data example: road data

- Roads are a common example of data that can be represented as a network
- Just having the physical shape of the network (polylines) is not sufficient
- A network requires **explicit connections** between road segments (= vertices)

There is no connection here (overpass)

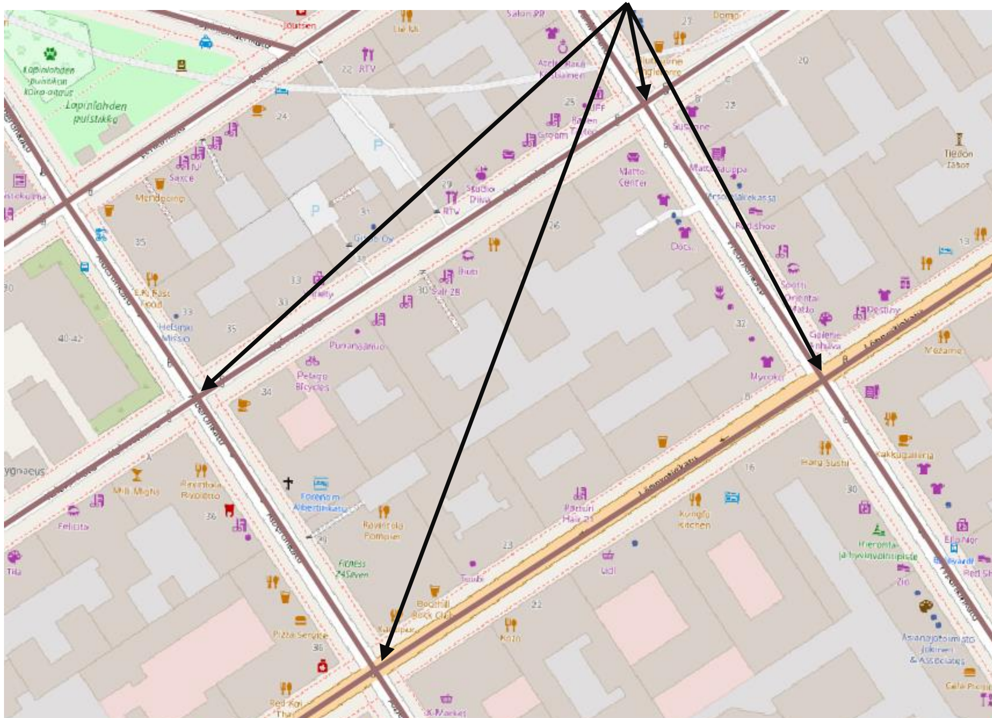


Aalto University
School of Engineering

Network data example: road data

- **Roads are a common example of data that can be represented as a network**
- **Just having the physical shape of the network (polylines) is not sufficient**
- **A network requires explicit connections between road segments (= vertices)**

Turns are restricted at these crossroads (one-way streets)



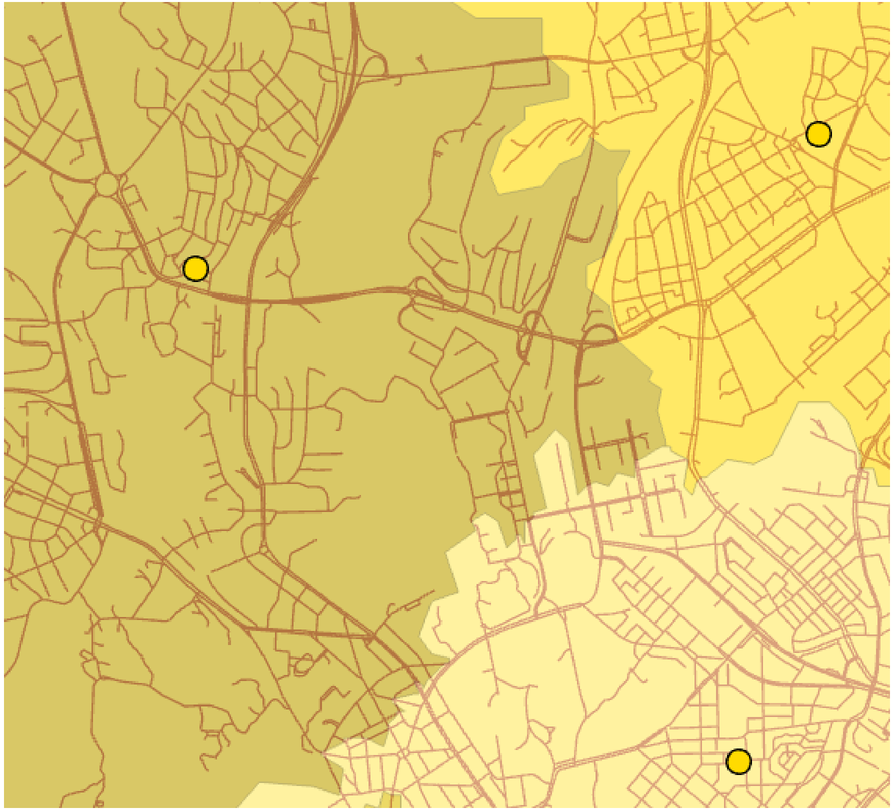
Network data example: road data

- **Road data is modeled as a graph**
- **Vertices are road junctions**
- **Edges are road segments between junctions**
  - One road typically consists of several segments
- **Divided highways are often represented by separate segments to each direction**

- **In road data elements need**
  - Direction (one- or two-way road)
  - Length, speed limit (can be used to calculate travel time)
  - Connections to other road data segments
  - Information about elements that can slow down movement (e.g. traffic lights)
  - Etc.

Network analysis example: service areas

- A graph can be used to solve the **shortest path problem**: what is shortest path from location A to other locations (or location B)
  - Reittiopas, google maps route, etc
- For a set of locations $p \in S$ we can use this to solve the **service area problem**: for all elements of the network, which location in S is closest?



Network analysis example: betweenness

- **The number of shortest paths a particular edge is part of defines the betweenness of the edge**
- **The higher the betweenness of an edge, the more important it is for the network**
  - ⟹ a vulnerable part of a road network

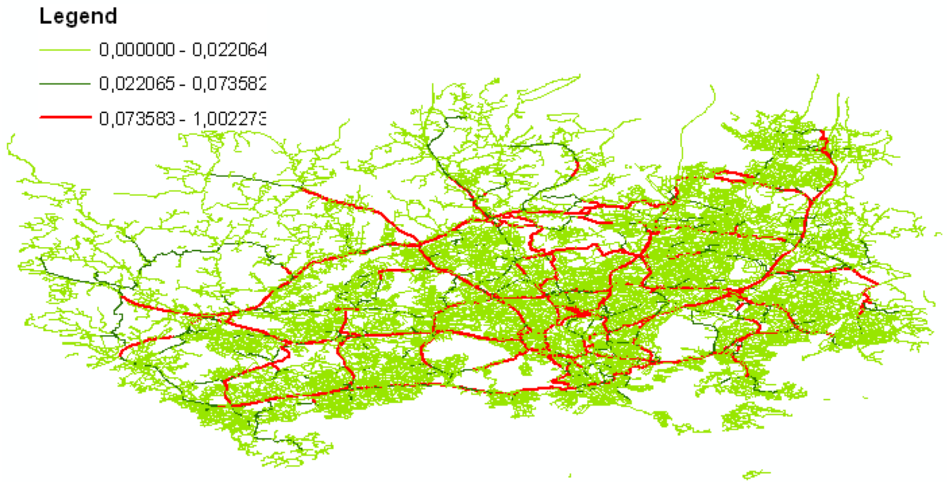


Image source: Zhang 2016

Network analysis example: betweenness

- **The number of shortest paths a particular edge is part of defines the betweenness of the edge**
- **The higher the betweenness of an edge, the more important it is for the network**
  - ⟹ a vulnerable part of a road network

Image source: Zhang 2016

Reading for the lecture

- **Longley et al. (2015): section 7.2.3.3 Network data model, section 7.2.3.4 TIN data model, chapter 13: spatial data analysis**
- **O’Sullivan and Unwin (2010): section 9.3 Spatial interpolation**