Learning over distributed chemical sensors data

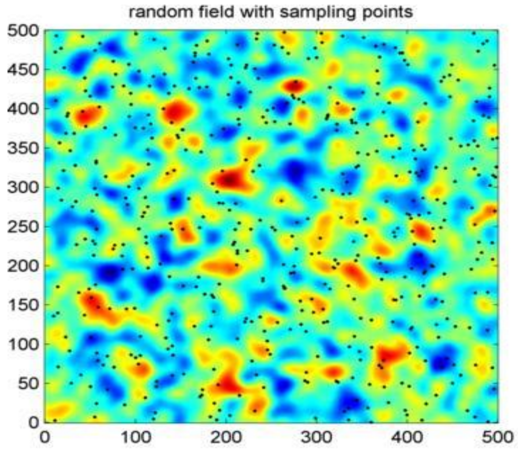
Santiago Marco

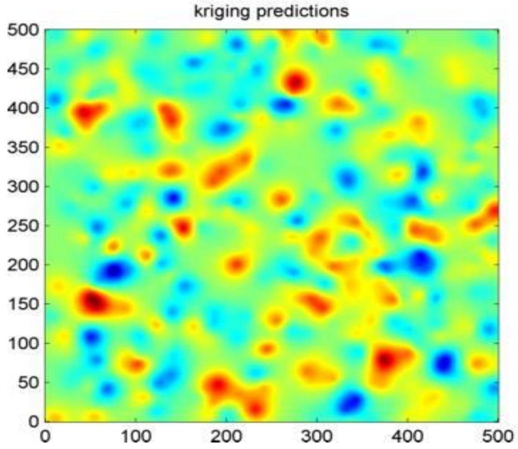
smarco@ibecbarcelona.eu

Signal and Information Processing for Sensing Systems Lab,

Institute for Bioengineering of Catalonia

Universitat de Barcelona





Outline:

- **Motivation:**
  - Availability of sensor nodes for distributed measurements of environmental variables.
  - Examples of hardware, application examples, data.
- **Basic concepts about learning spatial maps (Spatial interpolation)**
- **Review of methods**
  - Triangular Irregular Networks (TIN)
  - Inverse Distance Weighted (IDW)
  - Global and Local polynomial regression (LWR)
  - Radial Basis Functions (RBF)
  - Splines
  - Gaussian Processes / Kriging
- **Model evaluation**
- **Available tools**
- **Summary**

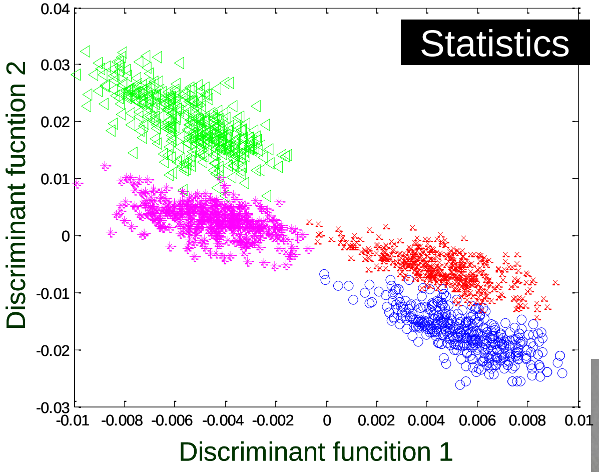
The Team

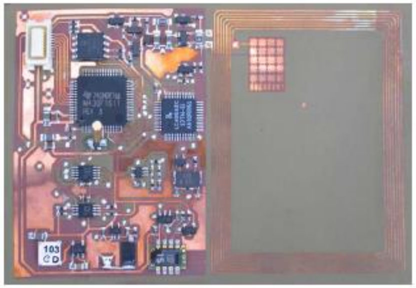
- IP: Santiago Marco, PhD in Physics
- Antonio Pardo, PhD in Physics
- Agustín Gutierrez, PhD in Computer Science

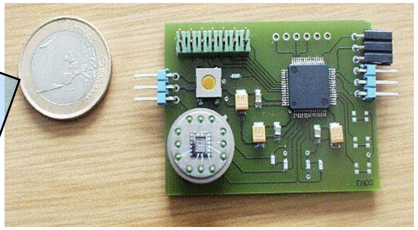
- **Post-doc**
  - *Jordi Fonollosa, PhD in Electronic Engineering*
  - *Marta Padilla, PhD in Electronic Engineering*
  - *Luis Fernández, PhD in Electronic Engineering*
  - *M. del Mar Contreras, PhD Analitical Chemistry*

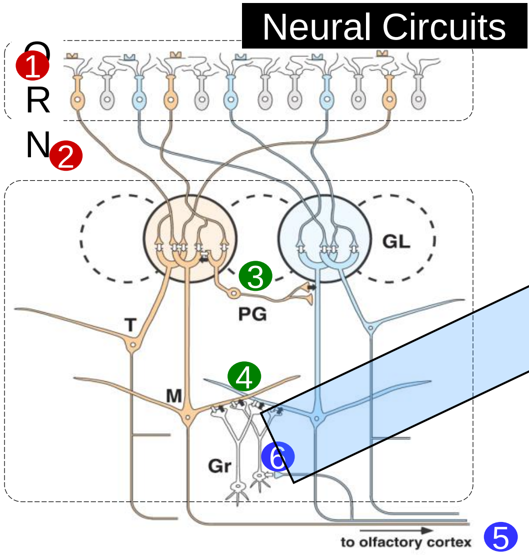
- **PhD Students**
  - *Sergi Oller, MSc Computational Physics*
  - *Ana Solorzano, MSc Electronic Engineering*
  - *Javier Burgues, MSc Computer Science,*

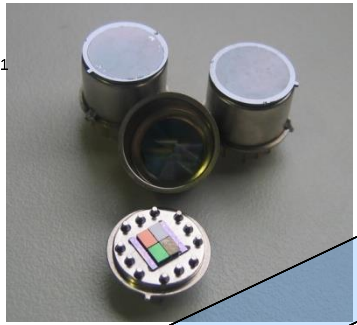
Mission

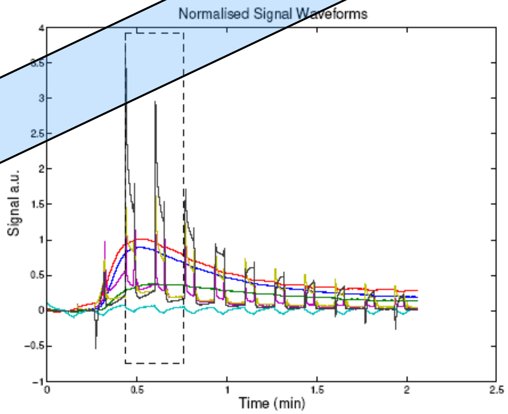












Mission:

Develop smart chemical sensor systems based on micro-nano technologies embedding advanced signal and data processing

DISTRIBUTED DATA

Distributed data

- **The issue of learning over distributed data is becoming more and more important due to the increasing availability of sensor nodes even for consumer applications.**
- Environmental surveys are almost always based on samples at **discrete points** (I will not cover particularly hyperspectral imaging) , but in general the measurements represent a **continuum in space** from which the sample has been drawn. Most analysts and their clients want to know what values are likely at intervening places

Italy
Rome
Naples
Tyrrhenian Sea
Adriatic Sea
Ionian Sea
Sicilia
Montenegro
Kosovo
Albania
Macedonia (FYROM)
Greece

https://www.netatmo.com/en-GB/weathermap

Distributed data

- **The number of profesional and consumer grade monitoring stations connected to the Internet of things are growing extremely fast**

Distributed data

- **A variety of companies sell OEM modules for chemical monitoring based on chemical sensors**

Libelium’s WASP mote

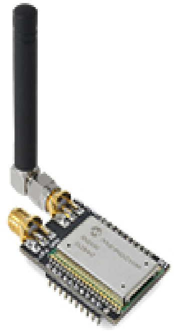
Distributed data: Wireless interfaces

» 802.15.4 / ZigBee

Model	Protocol	Frequency	Tx power	Sensitivity	Range*
XBee-802.15.4-Pro	802.15.4	2.4GHz	100mW	-100dBm	7000m
XBee-ZB-Pro	Zigbee-Pro	2.4GHz	50mW	-102dBm	7000m
XBee-868	RF	868MHz	315mW	-112dBm	12Km
XBee-900	RF	900MHz	50mW	-100dBm	10Km

* Line of sight and 5dBi dipole antenna.

» LoRaWAN 868 - 900/915 - 433MHz



Protocol: LoRaWAN 1.0, Class A

LoRaWAN - ready

Frequency: 868 MHz, 900 MHz and 433 MHz ISM frequency bands.

TX Power: up to +14 dBm

Sensitivity: as good as -136 dBm

Range: >15 km at suburban and >5 km at urban area. Typically, each base station covers some km. Check the LoRaWAN Network in your area.

» Sigfox

Range: Typically, each base station covers some km. Check the Sigfox Network

Max TX Power: 14dBm

Chipset: Telit LE51-868 S

Frequencies Available: 868-870MHz

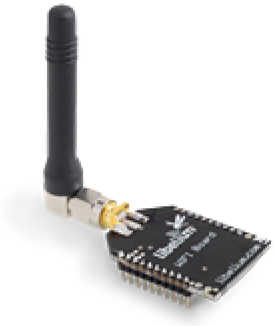
ETSI limitation: 140 messages of 12 bytes, per object per day

Chipset consuptiom: 55mA

Radio Data Rate: 100bps

Distributed data: Wireless interfaces

» WiFi



Protocols: 802.11b/g - 2.4GHz
TX Power: 0dBm - 12dBm (variable by software)
RX Sensitivity: -83dBm
Antenna connector: RPSMA
Antenna: 2dBi/5dBi antenna options
Security: WEP, WPA, WPA2
Topologies: AP & Adhoc
IP Setup: DHCP, Static

» Bluetooth Low Energy (BLE) 4.0

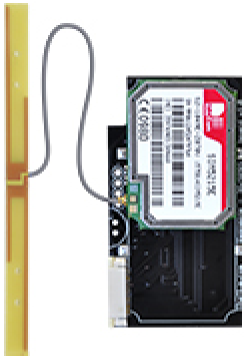
Protocol: Bluetooth v.4.0 / Bluetooth Smart
Chipset: BLE112
RX Sensitivity: -103dBm
TX Power: [-23dBm, +3dBm]
Antenna: 2dBi/5dBi antenna options
Security: AES 128
Range: 100 meters (at maximum TX power)
Consumption: sleep (0.4uA) / RX (8mA) / TX (36mA)

» 6LoWPAN / IPv6 Radio

6LoWPAN Radio (2.4GHz)

Chipset: AT86RF231
Frequency: 2.4GHz
Link Protocol: IEEE 802.15.4
Usage: Worldwide
Sensitivity: -101dBm
Security: WEP, WPA, WPA2
Output Power: 3dBm
Encryption: AES 128b

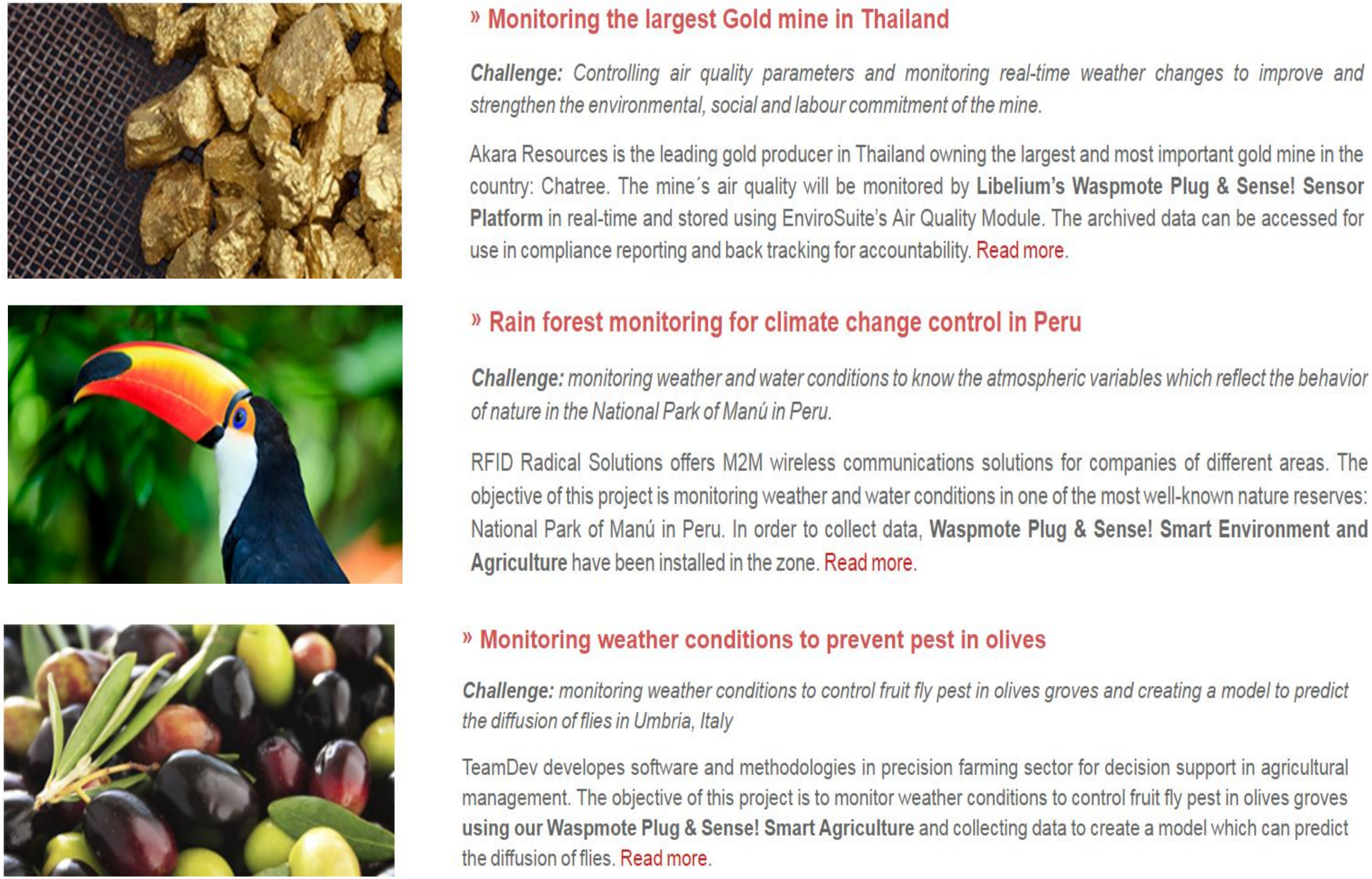
» 3G



Model: SIM5215
Protocols: 3G, WCDMA, UMTS, GPRS, GSM
Dual-Band: WCDMA/UMTS 900/2100 MHz
Tri-Band: UMTS 2100/1900/900MHz
WCDMA (downlink): up to 384Kbps
WCDMA (uplink): up to 384Kbps
TX Power:
UMTS 2100/900: 0.25 W
GSM 850/900: 2 W
DCS 1800: 1 W

Libelium catalog of products

Distributed data: Applications examples



» Monitoring the largest Gold mine in Thailand

Challenge: *Controlling air quality parameters and monitoring real-time weather changes to improve and strengthen the environmental, social and labour commitment of the mine.*

Akara Resources is the leading gold producer in Thailand owning the largest and most important gold mine in the country: Chatree. The mine´s air quality will be monitored by **Libelium's Waspmote Plug & Sense! Sensor Platform** in real-time and stored using EnviroSuite's Air Quality Module. The archived data can be accessed for use in compliance reporting and back tracking for accountability. Read more.

» Rain forest monitoring for climate change control in Peru

Challenge: *monitoring weather and water conditions to know the atmospheric variables which reflect the behavior of nature in the National Park of Manú in Peru.*

RFID Radical Solutions offers M2M wireless communications solutions for companies of different areas. The objective of this project is monitoring weather and water conditions in one of the most well-known nature reserves: National Park of Manú in Peru. In order to collect data, **Waspmote Plug & Sense! Smart Environment and Agriculture** have been installed in the zone. Read more.

» Monitoring weather conditions to prevent pest in olives

Challenge: *monitoring weather conditions to control fruit fly pest in olives groves and creating a model to predict the diffusion of flies in Umbria, Italy*

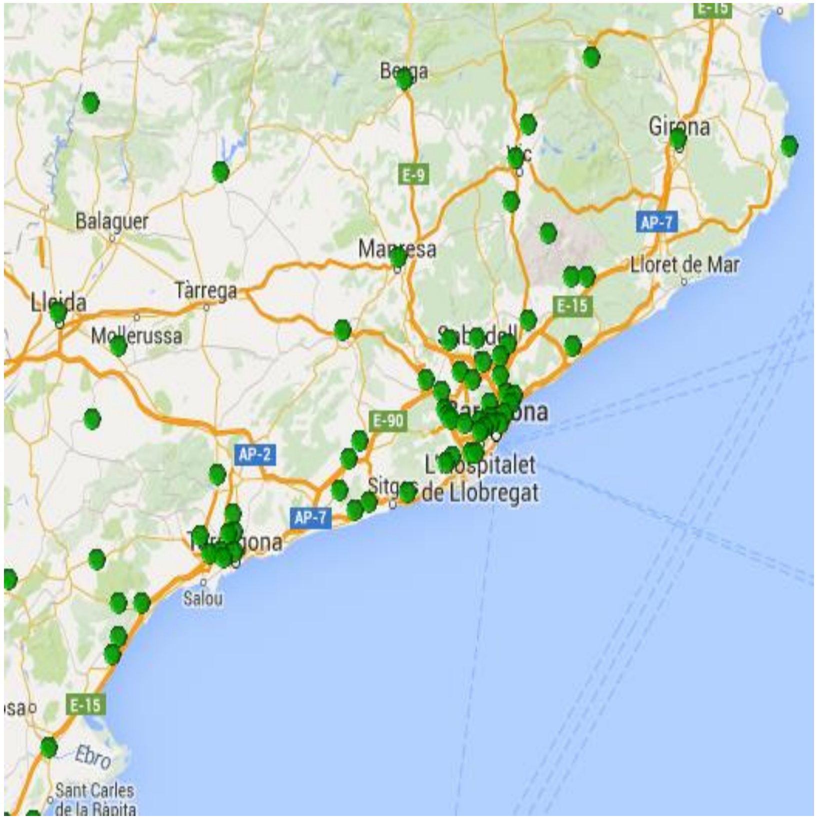
TeamDev developes software and methodologies in precision farming sector for decision support in agricultural management. The objective of this project is to monitor weather conditions to control fruit fly pest in olives groves **using our Waspmote Plug & Sense! Smart Agriculture** and collecting data to create a model which can predict the diffusion of flies. Read more.

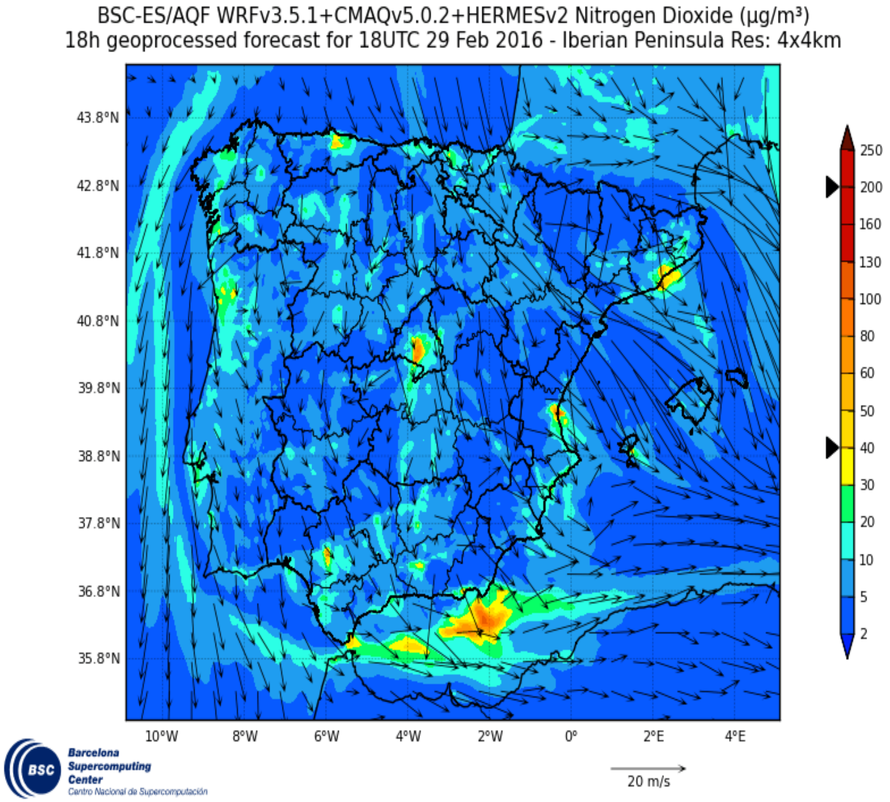
From libelium webpage: http://www.libelium.com

Distributed data: Pollution monitoring

- **Pollution monitoring is available only in a discrete network of pollution stations, but information is fused with metereological data to predict pollutant dispersión.**

http://www.bsc.es/caliope




BSC-ES/AQF WRFv3.5.1+CMAQv5.0.2+HERMESv2 Nitrogen Dioxide (µg/m³)
18h geoprocessed forecast for 18UTC 29 Feb 2016 - Iberian Peninsula Res: 4x4km
43.8°N
42.8°N
41.8°N
40.8°N
39.8°N
38.8°N
37.8°N
36.8°N
35.8°N
10°W
8°W
6°W
4°W
2°W
0°
2°E
4°E
250
200
160
130
100
80
60
50
40
30
20
10
5
2
20 m/s
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Spatially Distributed data

- **APPLICATION DOMAINS**
- **Environmental data is often collected at discret observation stations**
  - Environmental pollution
  - Meteorological station data

- **Geological data**
  - Soil type
  - Soil moisture
  - Vegetation data
  - Altitude (maps)

- **Main issues:**
  - Field data is expensive to collect or it can't be collected everywhere
  - Irregular sampling points
  - Partial data coverage
  - Irregular sampling times
  - Noisy observations
  - Interferences

Learning Goals

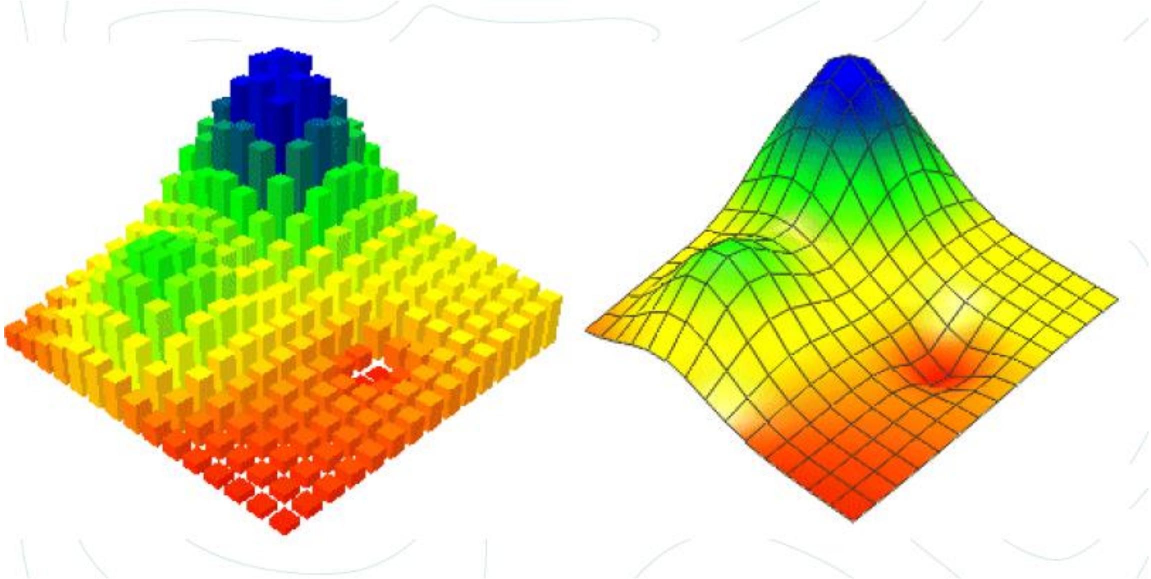
- Understand the need of learning over spatially distributed data
- Define spatial interpolation
- List global and local interpolation methods
- Differentiate between exact and approximate interpolation
- Describe kriging
- Understand the different sampling methods
- Get introduced on tools for spatial data management

- We will NOT describe throughly the maths of the different methods (just give some hints on methods behaviour) and basic equations.

BASIC CONCEPTS

Spatial Interpolation

- Features of the environment, such as pollution levels, are the product of ***many interacting physical, chemical and biological processes***. These processes are physically determined, but their ***interactions are so complex that the variation appears to be random***. This complexity and incomplete understanding of the processes means that a deterministic or mathematical solution to quantify the variation is out of reach at present.
- **Definition:**
  - “Spatial interpolation is the procedure of estimating the values of properties at unsampled sites within an area coverd by existing observations” (Waters, 1989).



Spatial Interpolation: Methods

Global All samples are used for the interpolation at a point	**Local** Only samples in the neighborhood of the point are used
Exact The map goes exactly through the measurement points	**Approximate** The map will pass close to the measurement points
Deterministic Rules are applied and the method provides the interpolation	**Stochastic** Samples are considered random variables

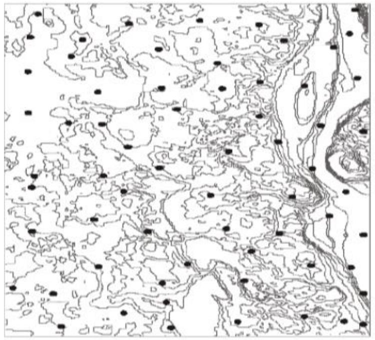
- **Triangular meshes: TIN (triangular irregular network)**
- **Inverse Distance Weighting (IDW)**
- **Global and Local Weighted Regression**
- **Regression Models: Radial Basis Functions (RBF)**
- **Splines**
- **Gaussian Processes and Kriging**

Spatial interpolation

- **Main underlying hypothesis:**
  - Interpolation is useful only if values are spatially dependent, or in other words if they are spatially autocorrelated.
  - Example of autocorrelated spatial data:
    - Temperature
    - Precipitation (rain))
    - Humidity
    - Pollutants distribution
- **The main hypothesis in the following presentation is:**
  - **Spatially close mearurements are more correlated than distant measurements**.

Spatial Interpolation

Sampling Design
Data Collection
Model Building



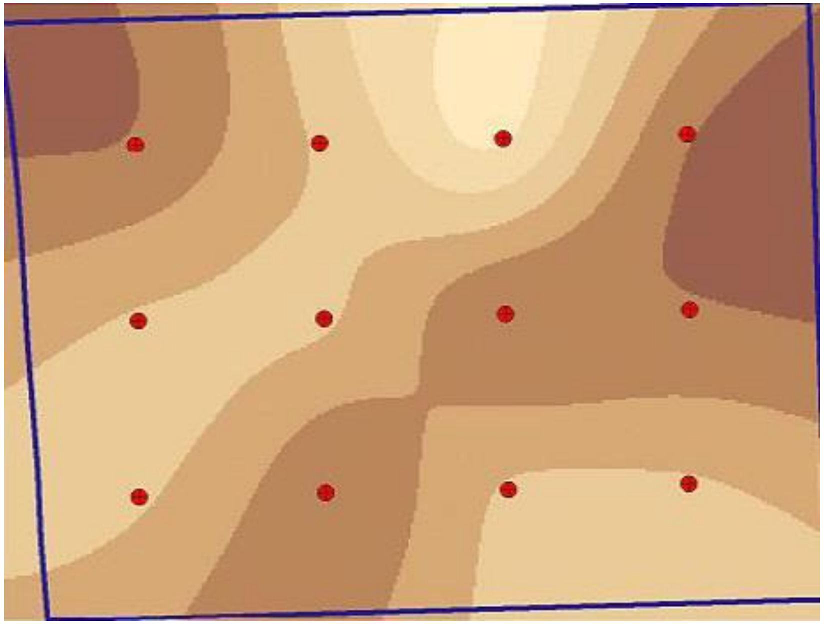
$$Z(x,y) = \sum_{k=1}^{M} a_k G_k(x,y)$$

Spatial interpolation

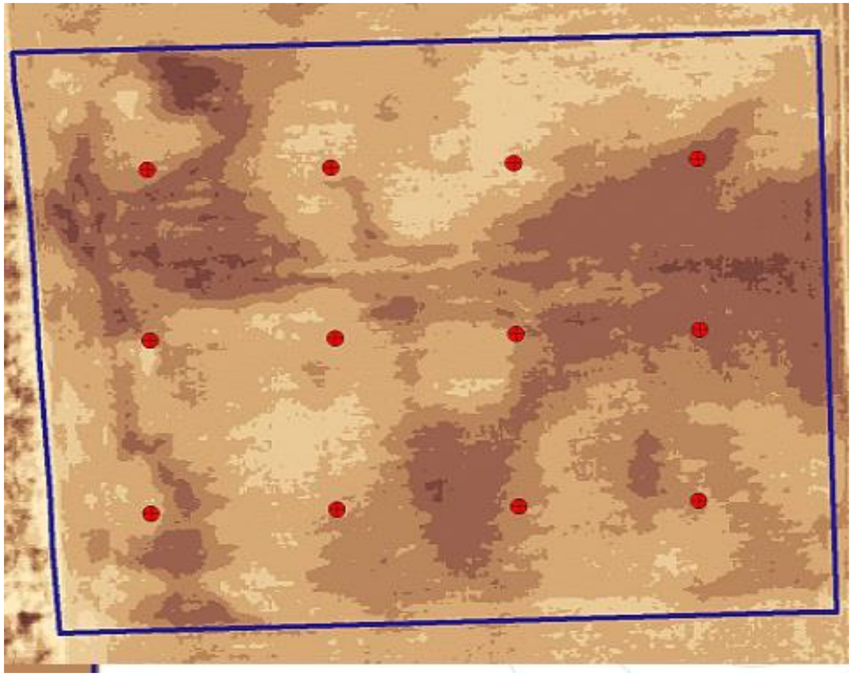
- **Data types:**
  - **Dense** data typically from hyperspectral imaging.
  - **Sparse data** originating from discrete monitoring stations.

- **Example: Percentage of covering vegetation.**

J.P. Mund, 2013

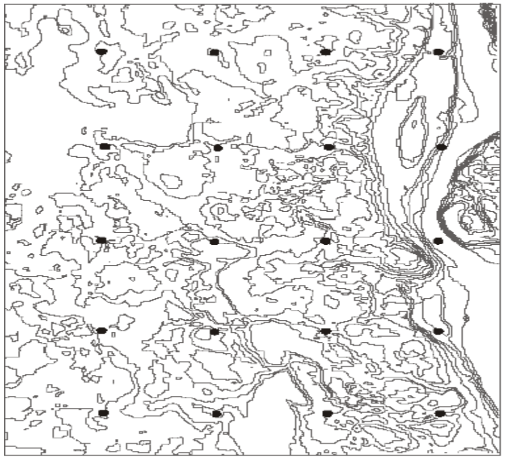


Interpolated from discrete measurements

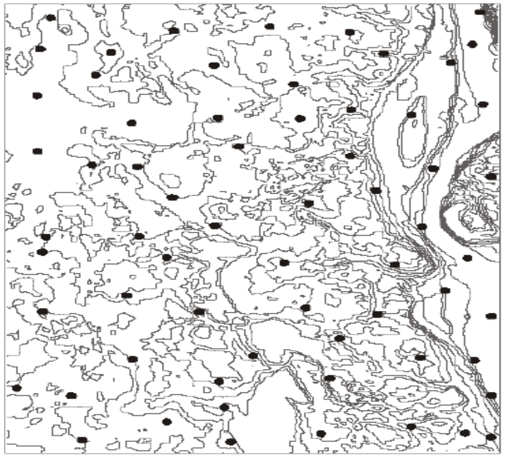


Aerial measurements in IR band

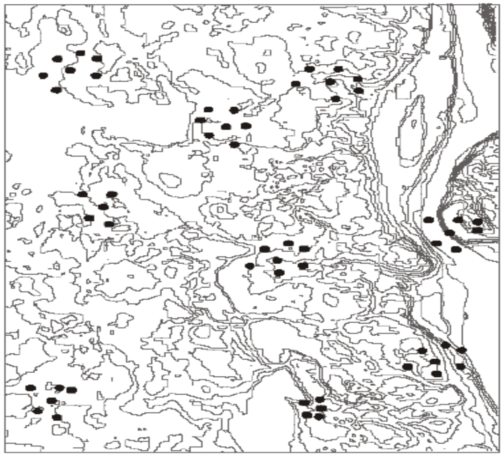
Spatial Interpolation: Data sampling



Uniform sampling

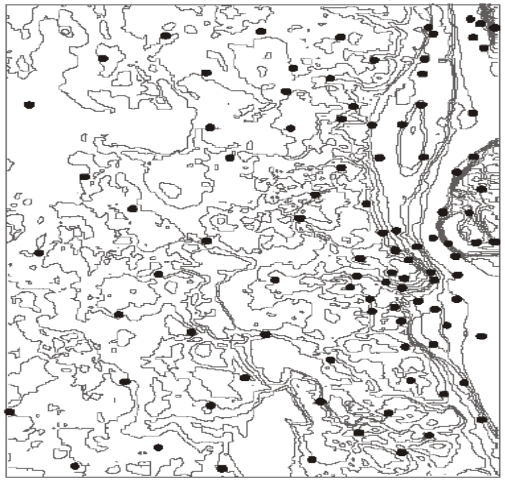


Random Sampling



Cluster Sampling

- Uniform sampling is not always posible due to practical difficulties and produces bias
- Cluster sampling provides measurements in most accesible áreas (e.g. towns)
- Adaptive sampling proposes higher density poitns in areas of higher variability, it is typically an interative process.



Spatial Interpolation

- **Exact Interpolation:**

From $\{x_i, y_i, z_i\}\ i = 1 \ldots N$ find $\mathrm{Z}:\mathbb{R}^2 \to \mathbb{R}$ such that $\mathrm{Z}(x_i, y_i) = z_i$

Using Generalized Linear Models (GLM) exact interpolation results in a system of equations with N points and N unknown coefficients. The choice of the basis functions is critical.

$$\mathrm{Z}(x, y) = \sum_{k=1}^{N} a_k G_k(x, y)$$

The problem of finding the interpolation coefficients can be stated as a matrix equation:

$$\begin{pmatrix} G_1(x_1, y_1) & \cdots & G_N(x_1, y_1) \\ \vdots & \ddots & \vdots \\ G_1(x_N, y_N) & \cdots & G_N(x_1, y_1) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix} \quad \text{or} \quad \boldsymbol{Ga} = \boldsymbol{z}$$

Provided G is non singular $\boldsymbol{a} = \boldsymbol{G}^{-1}\boldsymbol{z}$

Spatial Intepolation

Approximate interpolation:

- When experimental points have a substantial amount of noise, regression is more robust than interpolation.

From $\{x_i, y_i, z_i\}\ i = 1 \ldots N$ find $\mathrm{Z}{:}\mathbb{R}^2 \to \mathbb{R}$ such that $\mathrm{Z}(x_i, y_i) \sim z_i$

Typically the approximation is enforced in a least square sense.

Using Generalized Linear Models (GLM) exact interpolation results in a system of equations with N points and M<<<N unknown coefficients.

$$\mathrm{Z}(x, y) = \sum_{k=1}^{M} a_k G_k(x, y)$$

The problem of finding the interpolation coefficients can be stated as a matrix equation:

$$\begin{pmatrix} G_1(x_1, y_1) & \cdots & G_M(x_1, y_1) \\ \vdots & \ddots & \vdots \\ G_1(x_N, y_N) & \cdots & G_N(x_N, y_N) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_M \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_N \end{pmatrix} \quad \text{or} \quad \boldsymbol{Ga} = \boldsymbol{z}$$

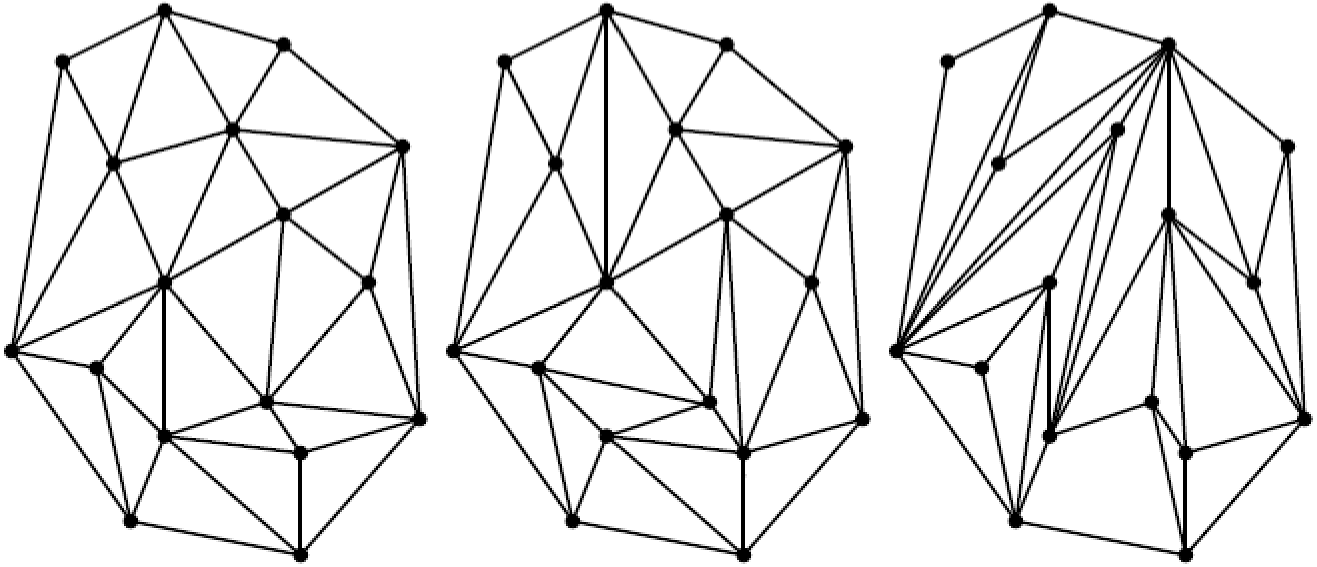
Provided G is well conditioned $\boldsymbol{a} = \boldsymbol{G}^{+}\boldsymbol{z}$ where $\boldsymbol{G}^{+} = \left(\boldsymbol{G}^{T}\boldsymbol{G}\right)^{-1}\boldsymbol{G}^{T}$

To ensure particular behaviour on the solution additional objective functions with penalty functions. This is a regularization approach.

SPATIAL INTERPOLATION METHODS

Exact Spatial Interpolation: TIN

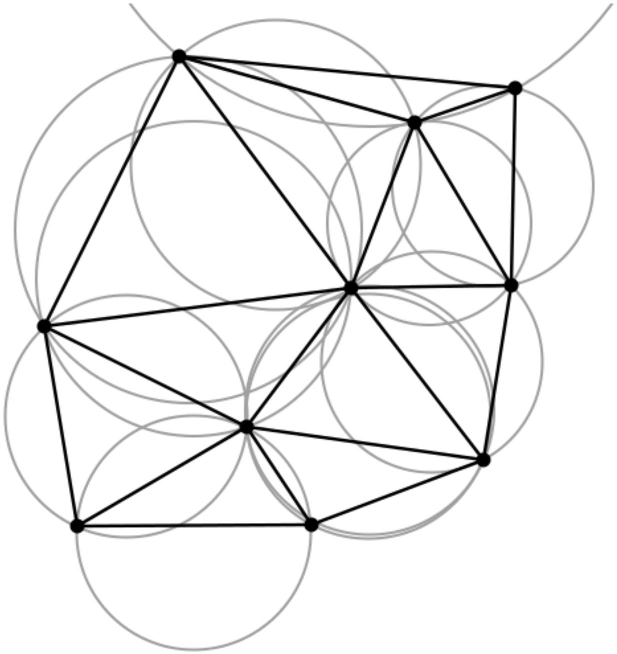
- **Triangular meshes (Peuker -1978): Triangular Irregular networks**
  - This exact interpolation starts by definin a triangular mesh on the {x,y} map.
  - There could be many triangular meshes:



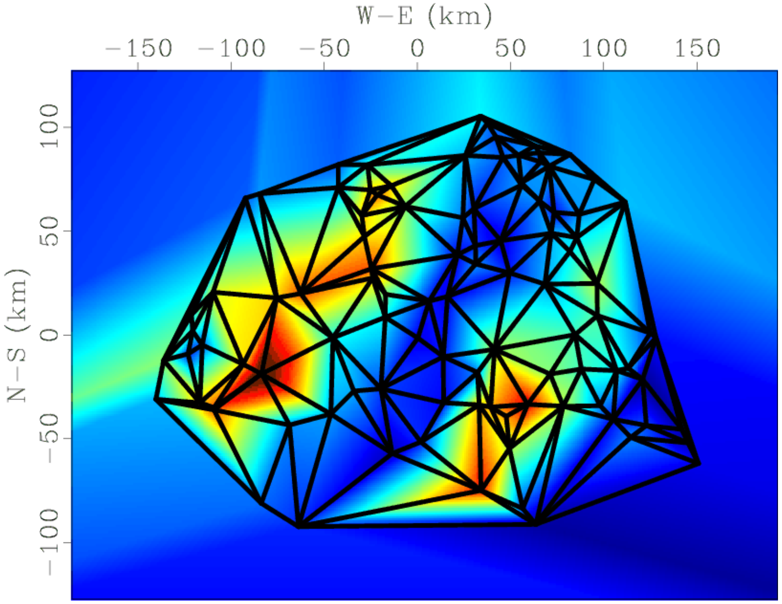
For interpolation, triangles should not be with small angles.

Exact Spatial Interpolation: TIN

- **Triangular meshes: Delaunay Triangulation**
  - A Delaunay triangulation for a set of P points in a plane is a triangular mesh such that no point in P is inside the circumference defined by any triangle in the mesh.



Within each triangle values are interpolated by linear interpolation or cubic polynomial interpolation (Ripley, 1981)

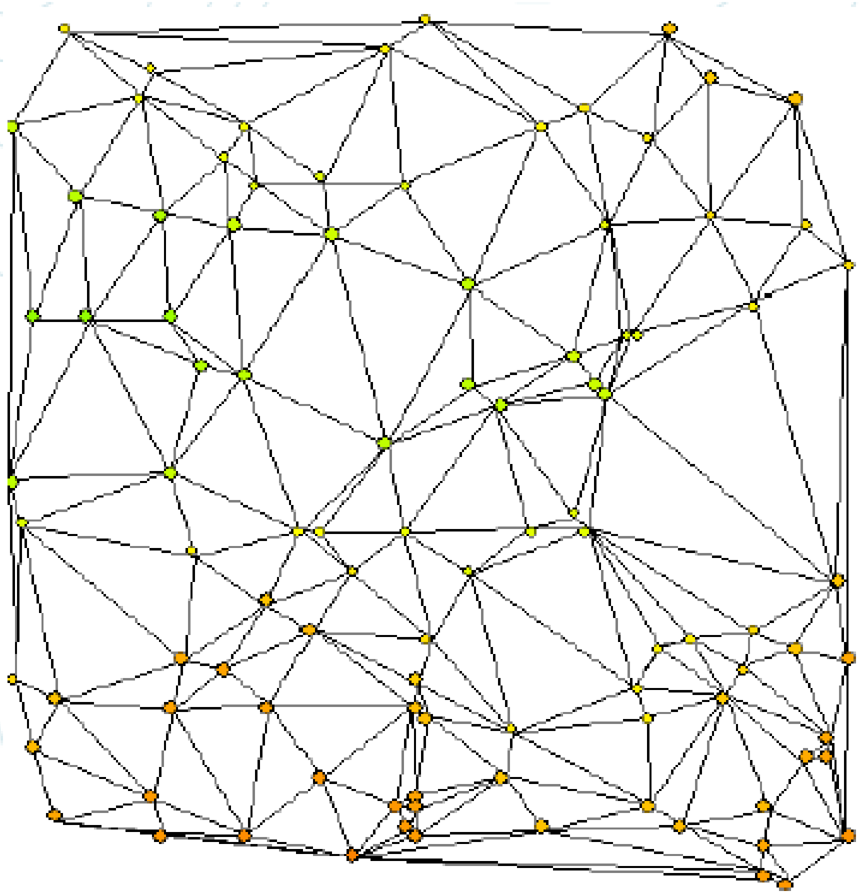


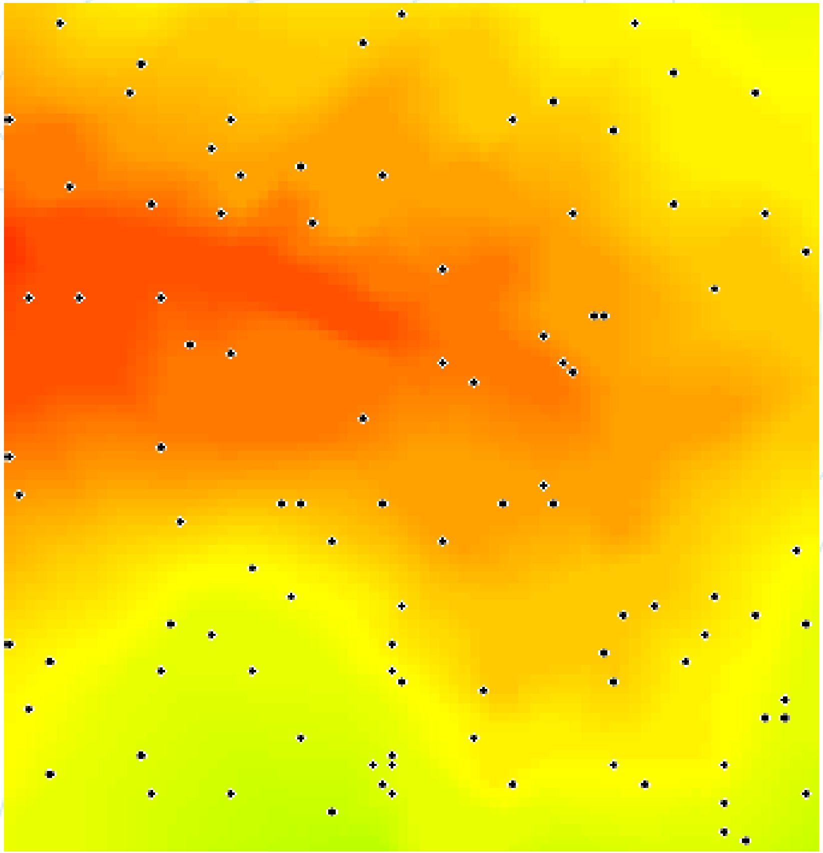
Delaunay Triangulation

Image from wikipedia

Exact Spatial Interpolation: TIN

- **Dulanay triangularization and cubic interpolation**





J.P. Mund, 2013

Exact Spatial Interpolation: IDW

- **Inverse Distance Weighted (IDW)**

- **Each point has a local influence that diminishes with distance.**

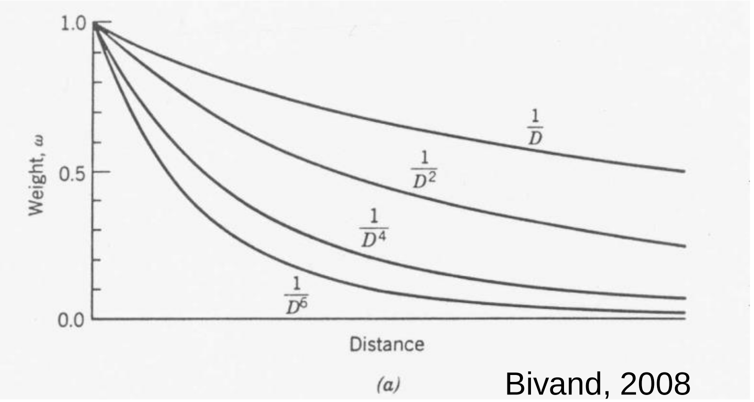
$$z(x) = \frac{\sum_i w_i z_i}{\sum_i w_i}$$

$$w_i = {}^{1}/_{d_i^n}$$

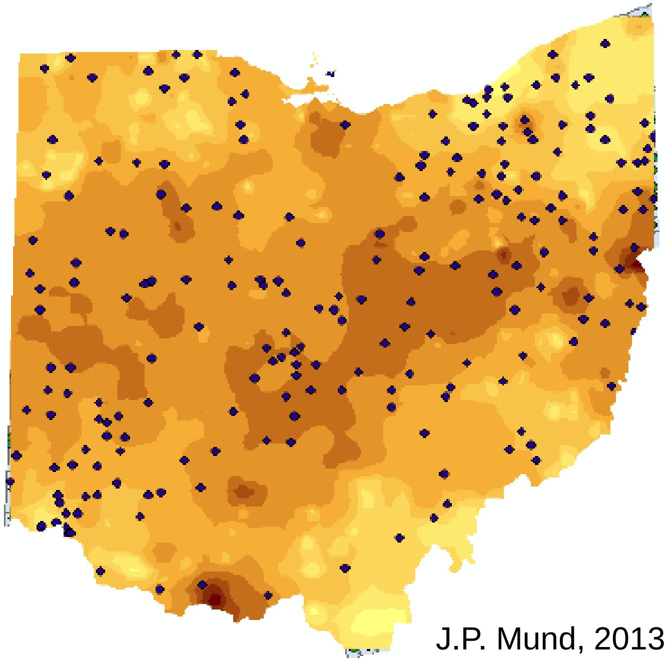
n is typically 1,2,3

Sum is over all points

- **Undesired characteristic of IDW:**
  - In areas with no data far away from the simple pointts IDW just tends to the mean of data.



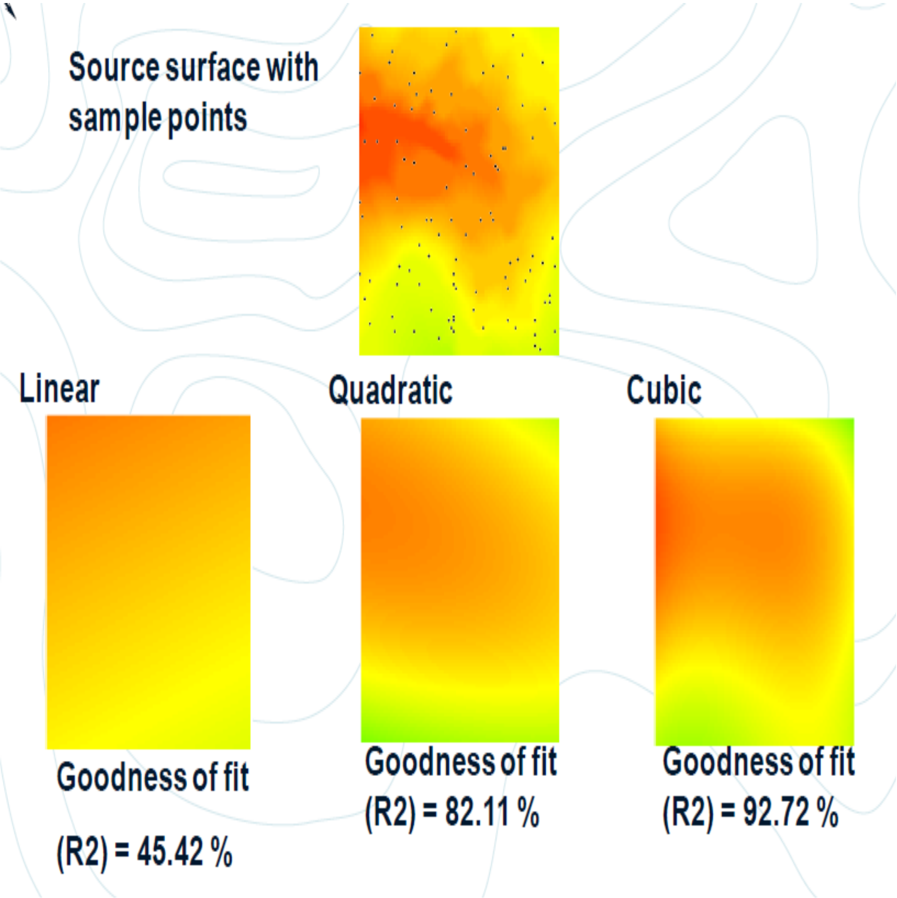
Bivand, 2008

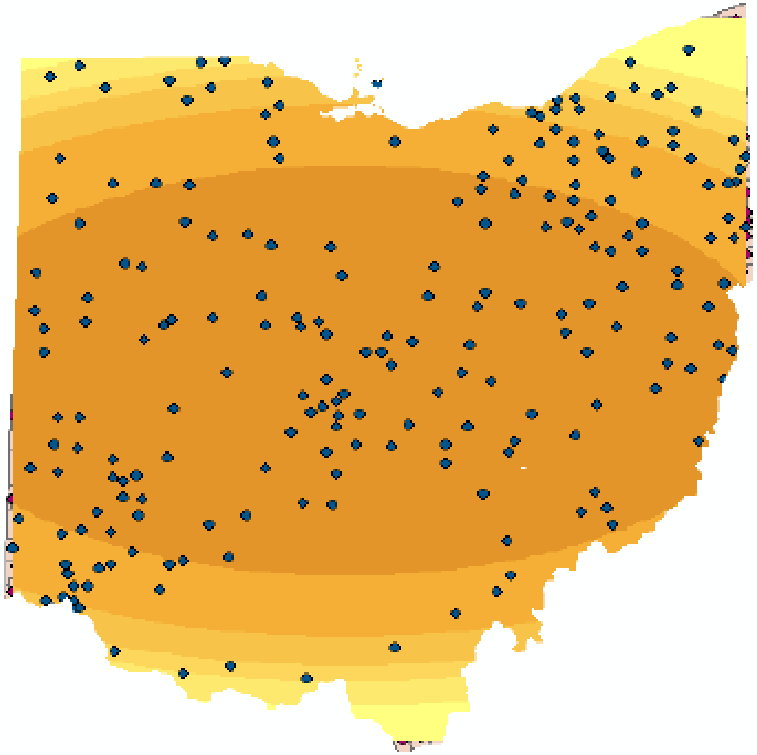


J.P. Mund, 2013

Approximate Spatial interpolation: Global

- **Whole area polynomial interpolators (trend estimation):**
  - Sometimes beyond local features the local trend is estimated with polinomials over the whole exploration área to extract smooth trends.

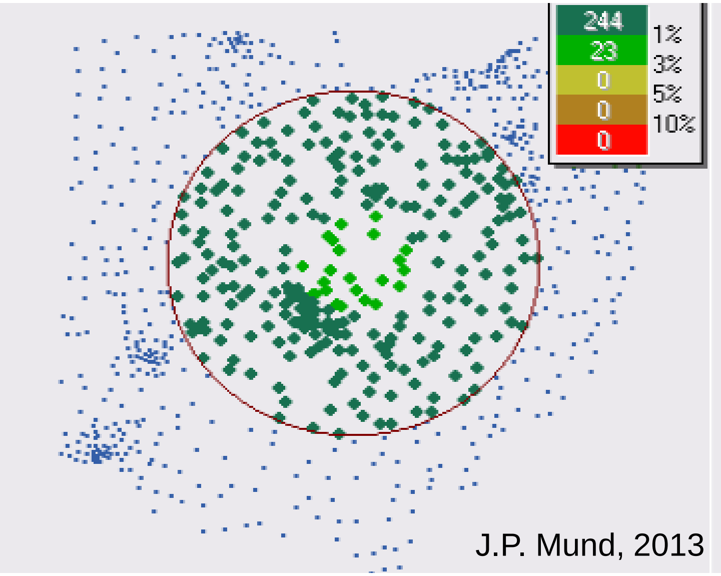




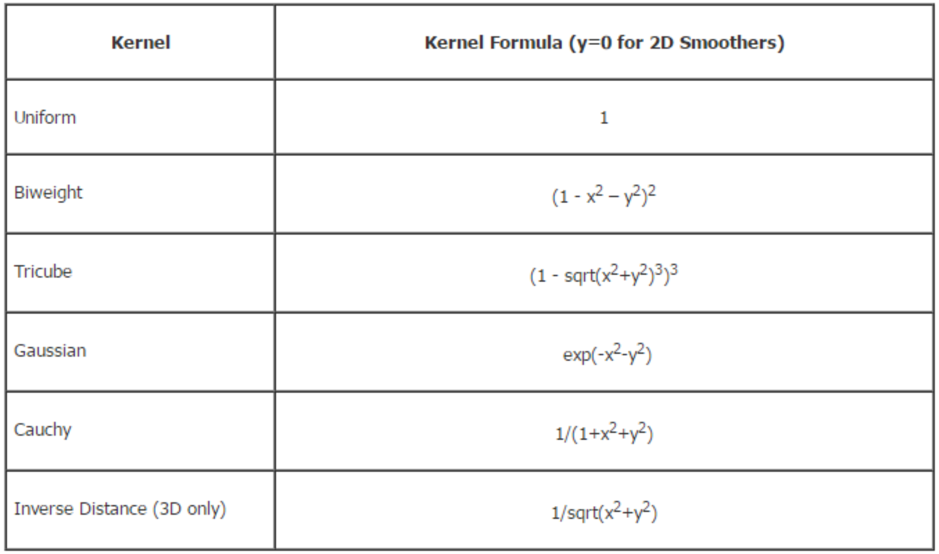
J.P. Mund, 2013

Approximate Spatial interpolation: Local

- **Local Polynomial interpolation fits polynomials based on neighrest neighbours to point of interests (Local Weighted Regression)**
  - Neighborhood can be defined in terms of size and shape, or in terms of numbers of neighbours
  - Typically polinomial orders vary between 1 and 3.
  - Typically weighted leasts squares is used for model estimation where closer points are given more weight than distant points.
  - A number of 'kernels' can be used to define the dependence on distance.
  - Even anistotropic kernels have been proposed.



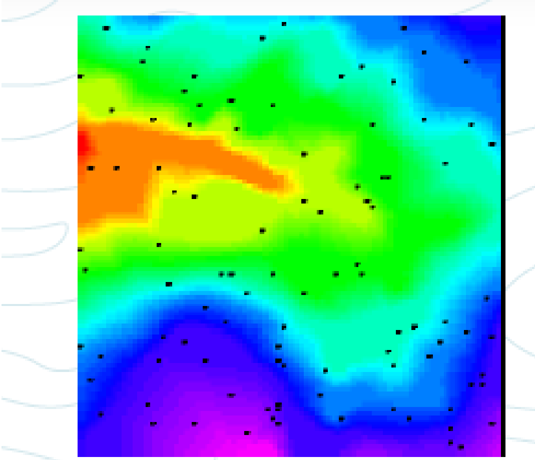
J.P. Mund, 2013



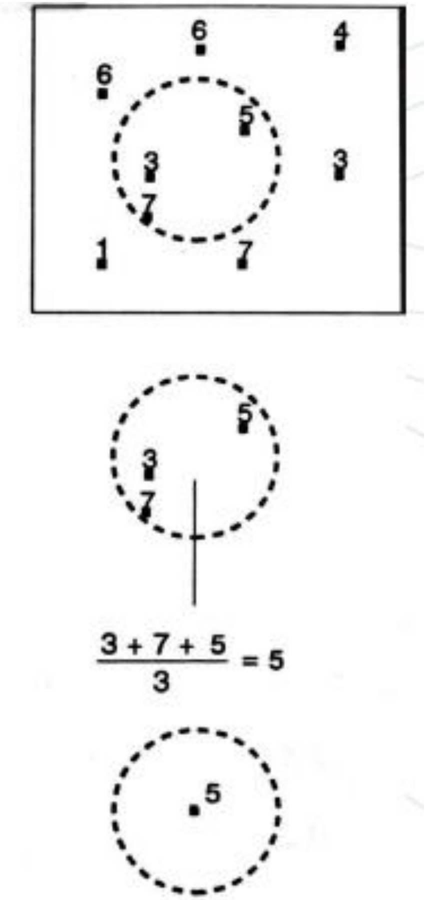
Kernel	Kernel Formula (y=0 for 2D Smoothers)
Uniform	1
Biweight	$(1 - x^2 - y^2)^2$
Tricube	$(1 - \mathrm{sqrt}(x^2+y^2)^3)^3$
Gaussian	$\exp(-x^2-y^2)$
Cauchy	$1/(1+x^2+y^2)$
Inverse Distance (3D only)	$1/\mathrm{sqrt}(x^2+y^2)$

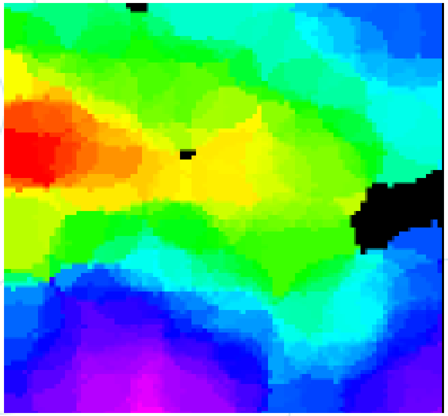
Approximate Spatial Interpolation

- **SMA: Spatial Moving Average.**
  - Simplified form of LWR (no weights, polynomial order zero)

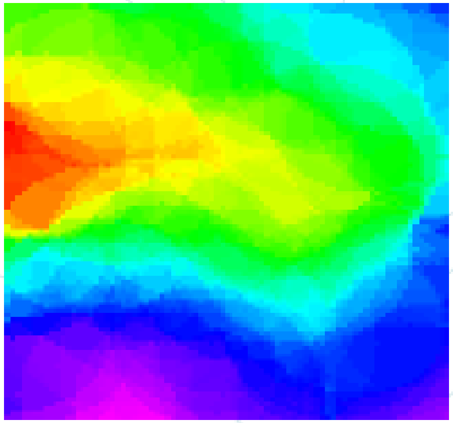


Source surface with sample points





21x21 circular filter SMA



41x41 circular filter SMA

Radial Basis Function Networks

Exact interpolation (Powell, 1987):

Data points

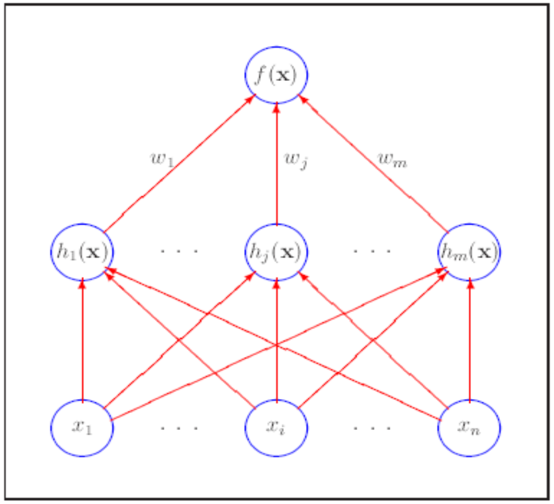
$$h(x) = \sum_n w_n \Phi\left(\left\| x - x^n \right\| \right)$$

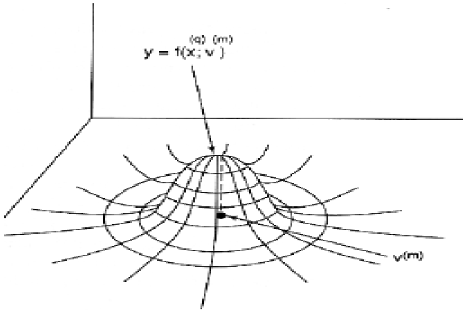
$$\Phi(x) = \exp\left(-\frac{x^2}{2\sigma^2} \right)$$

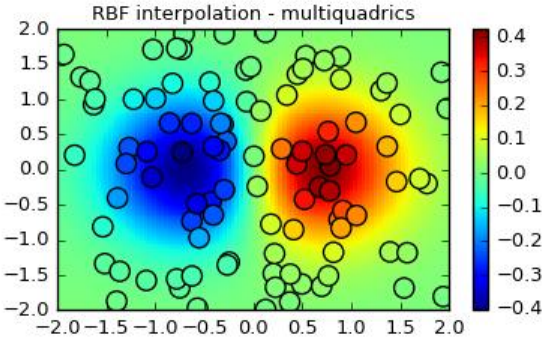
A regressor centered in every data point.

The weight parameters may be obtained by matrix inversion:

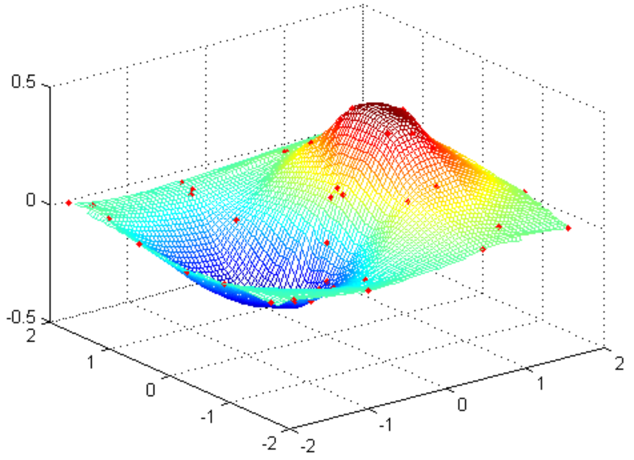
$$W = \Phi^{-1} t$$

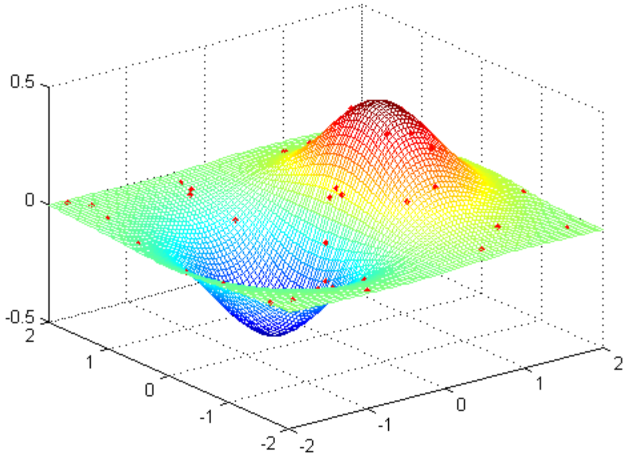


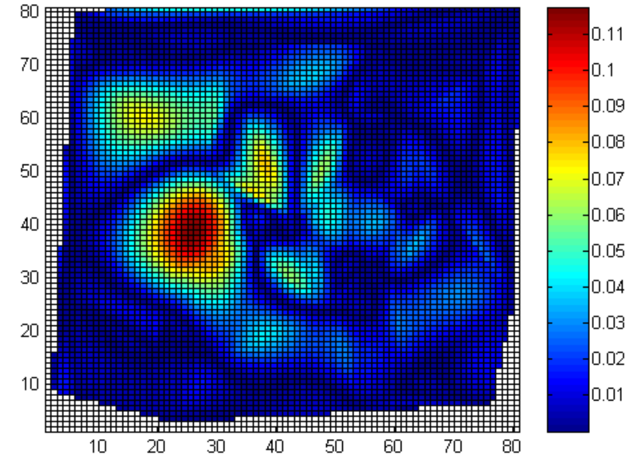


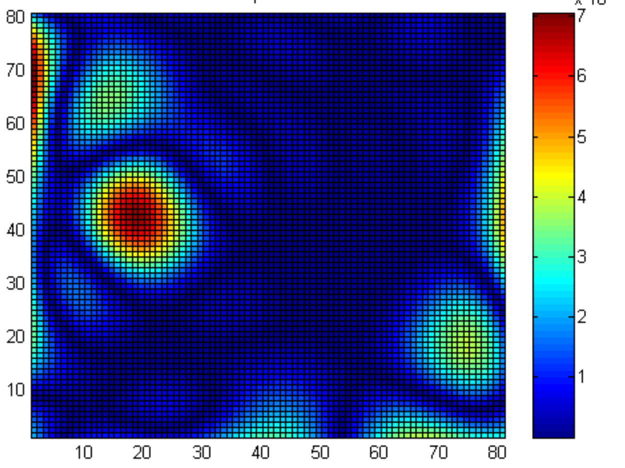


Approximate Spatial Intepolation: RBF


Interpolation using Matlab function griddata(method=cubic)

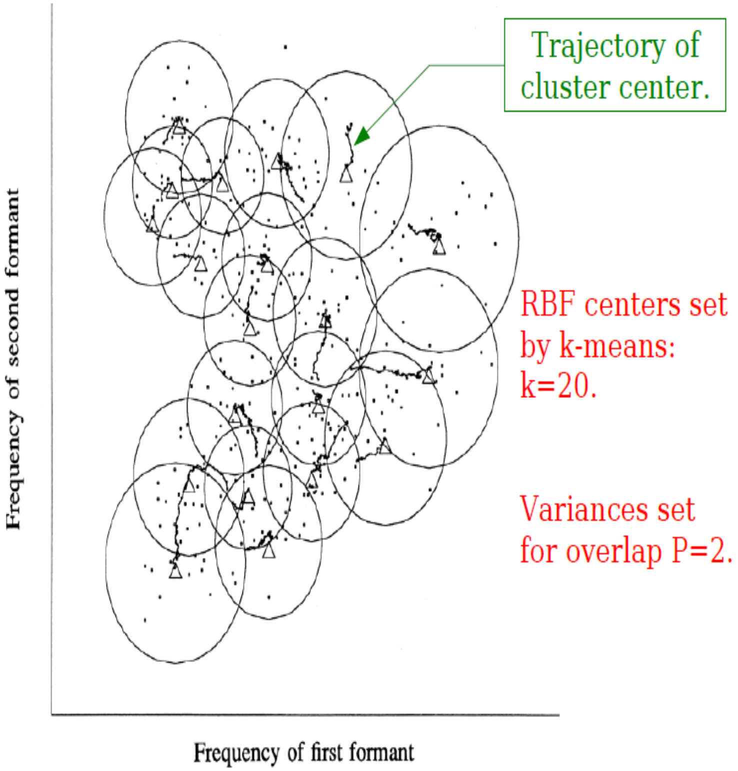

RBF interpolation


griddata(method=cubic) interpolation error


RBF interpolation error

Radial Basis Function Networks

- **Modifications from the exact interpolation approach towards approximate interpolation.**
  - The number M of basis functions need not to equal the number of data points and is typically much less than N
  - The centres of the basis fucntions are no longer constrained to be given by the input data vectors
  - The determination of the centers becomes part of the training process
  - Each basis function may have its own covariance
- **Once centers have been chosen second layer can be trained by linear least squares.**



$$W^T = \Phi^+ T$$

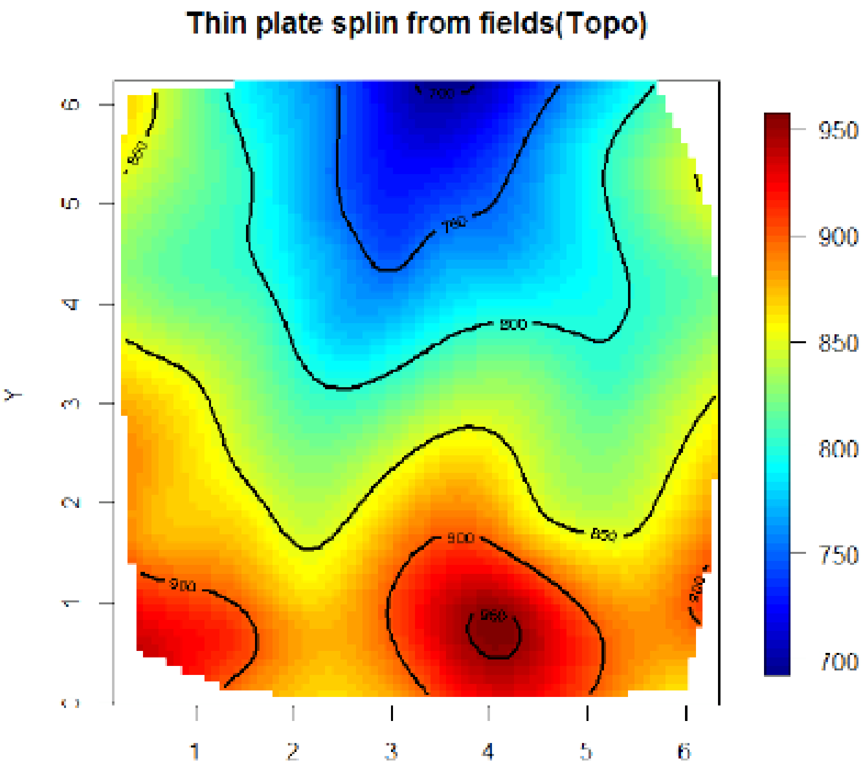
$$\Phi^+ \equiv \left(\Phi^T \Phi\right)^{-1} \Phi^T$$

Spline Regression: Thin Plate Splines

- **The thin plate spline uses a kernel of the form:**

- $\boldsymbol{u(r, r_i) = |r - r_i|^2 \, ln \left(\frac{|r - r_i|}{d}\right)}$

- **Given a set of data points, a weighted combination of thin plate splines centered on the data points provides an exact interpolation while minimizing the total curvature (bending energy).**



$$\hat{Z}(x, y) = a_0 + a_1 x + a_2 y + \sum_{i=1}^{N} b_i \, u(r, r_i)$$

Laplacian Smoothing Spline (Thin Plate Splines)

- **The smoothing spline is a method of fitting a smooth curve to a set of noisy observations using a spline function.**
- **The smoothing spline minimizes the following objective function over the class of twice diferenciable functions.**

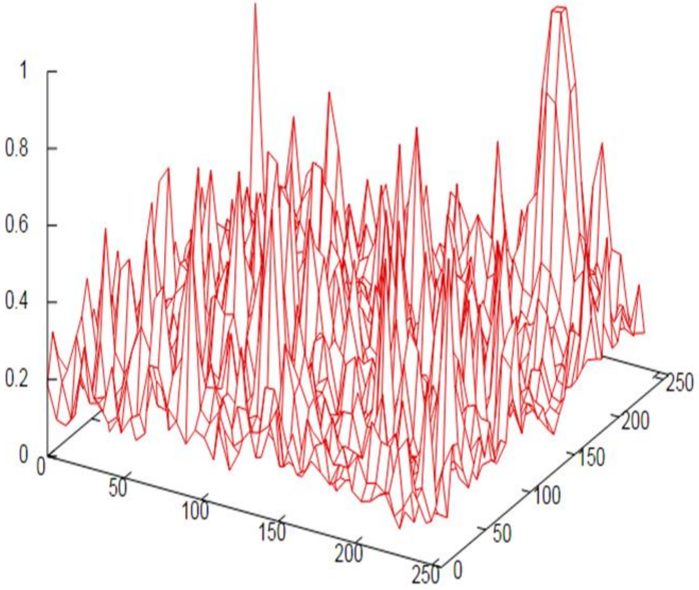
$$Q = \sum_{i=1}^{N} \left(z_i - Z(x_i, y_i)\right)^2 + \lambda J_2(Z)$$

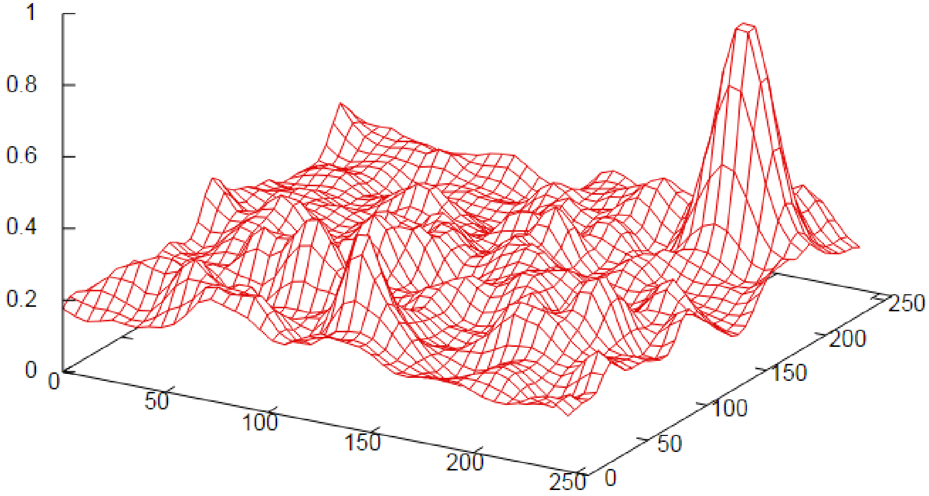
$$J_2(Z) = \iint dxdy \left(Z_{xx}^2 + 2Z_{xy}^2 + Z_{yy}^2\right)$$

$$Z_{xx}^2 \equiv \frac{\partial^2 Z(x,y)}{\partial x^2}$$

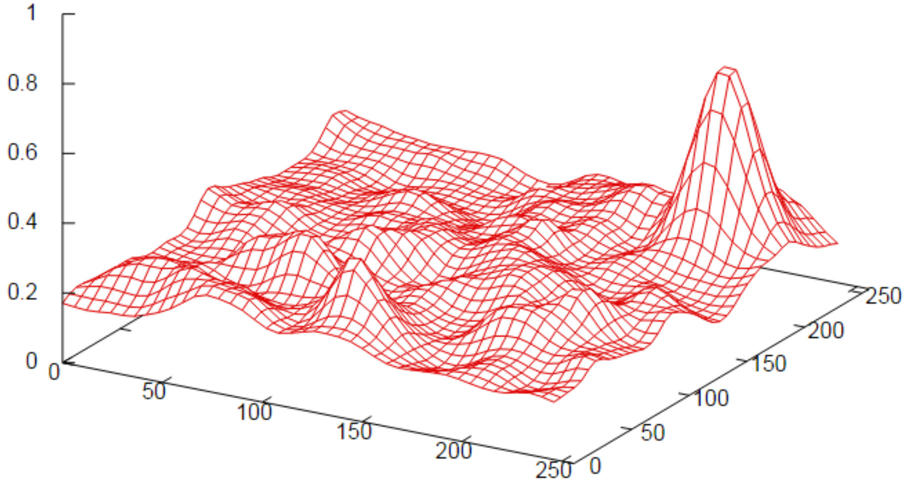
The regularization parameter is found by CV.

Spline interpolation in noisy data





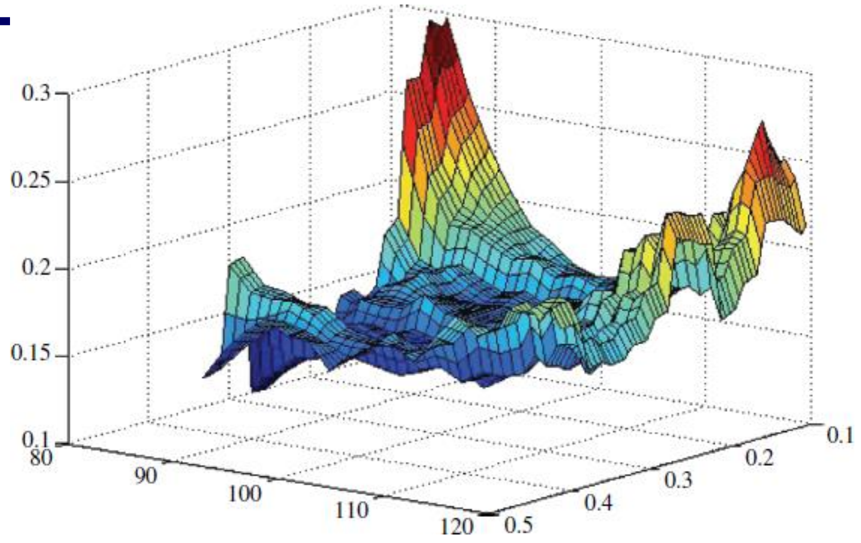
Exact interpolation
Approximated interpolation
with two values of the
regularization parameter

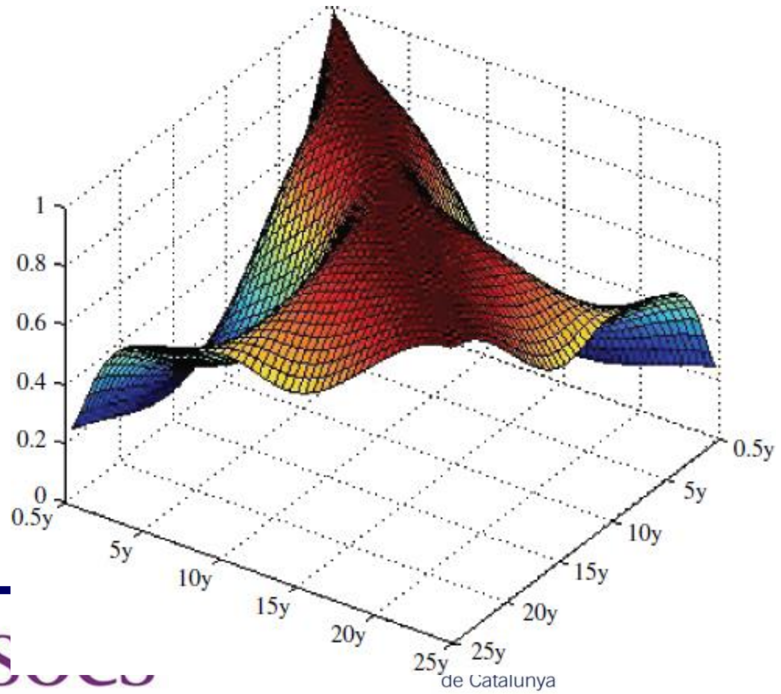


Random multivariate signals

- The value of a property, say z at any place (x,y) denoted z(x,y) is just one of the many infinite values of the random variable at that place. Its a 'realization' of the process.
- When we consider the infinite collection of random variables we say is a random process.
- We hypothesize that closer random variables are more correlated.
- More correlation can lead to smoother maps.

- Problem; at a given time we have only a given realization.

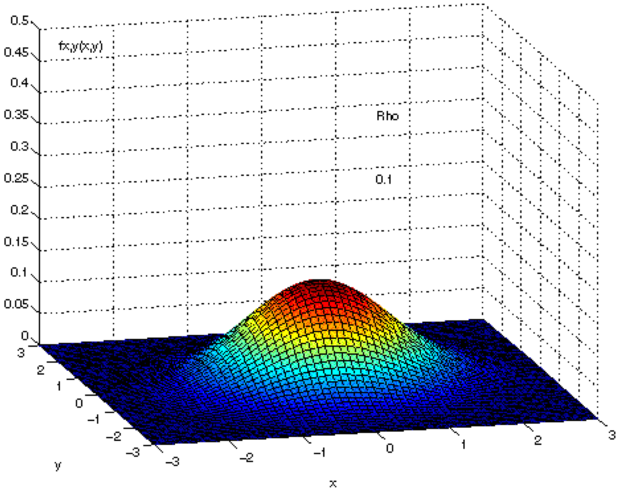




Review of Gaussian Random Variables

- **Meet the multivariate Normal or Gaussian density N(μ,Σ):**

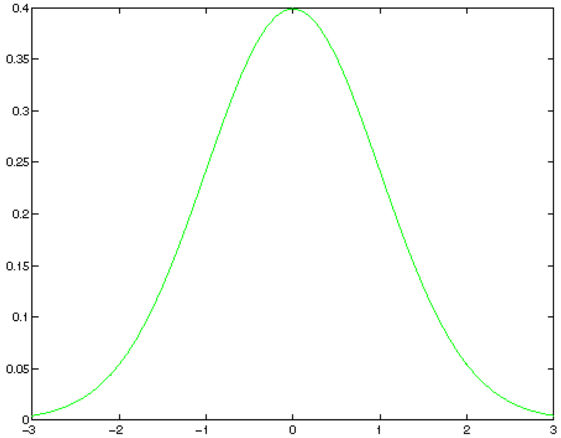
$$f_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right]$$



  - For a single dimension, this expression reduces to the familiar expression

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

- **Gaussian distributions are very popular**
  - The parameters (μ,Σ) are **sufficient** to uniquely characterize the normal distribution
  - If the $\mathbf{x_i}$'s are mutually **uncorrelated** ($\mathbf{c_{ik}}$=0), then they are also **independent**
    - The covariance matrix becomes diagonal, with the individual variances in the main diagonal
  - Central Limit Theorem
  - Marginal and conditional densities
  - Linear transformations



Materials in this slide adapted from Dr. Ricardo Gutierrez, Texas A&M, USA

Review of probability theory

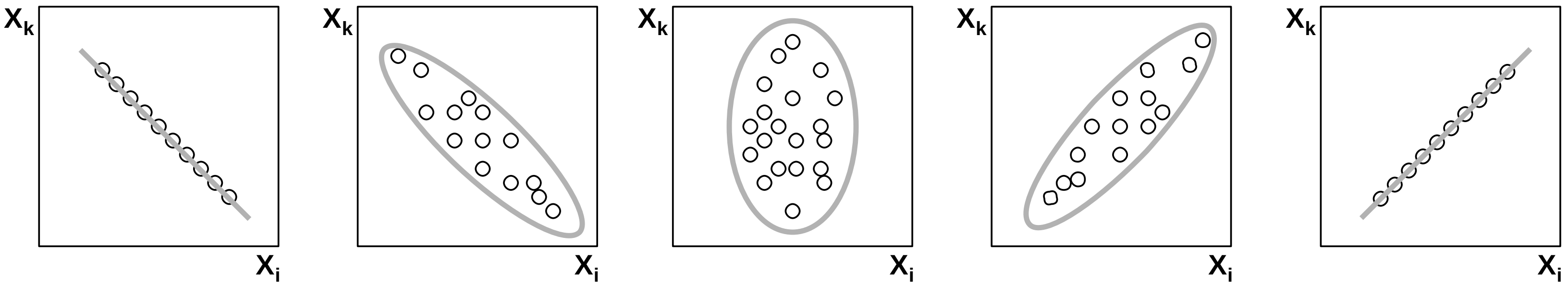
- **The covariance matrix (cont.)**

$$\mathrm{COV}[X] = \Sigma = E[(X-\mu)^T(X-\mu)] \approx \frac{1}{N}(X-\mu)^T(X-\mu)$$

$$= \begin{bmatrix} E[(x_1-\mu_1)(x_1-\mu_1)] & \dots & E[(x_1-\mu_1)(x_N-\mu_N)] \\ \dots & \ddots & \\ E[(x_N-\mu_N)(x_1-\mu_1)] & \dots & E[(x_N-\mu_N)(x_N-\mu_N)] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \dots & c_{1N} \\ \dots & \dots & \\ c_{1N} & \dots & \sigma_N^2 \end{bmatrix}$$

- The covariance terms can be expressed as $c_{ii} = \sigma_i^2$ and $c_{ik} = \rho_{ik}\sigma_i\sigma_k$
  - where ρ_{ik} is called the correlation coefficient

- **Graphical interpretation**



$C_{ik}=-\sigma_i\sigma_k$
$\rho_{ik}=-1$

$C_{ik}=-½\sigma_i\sigma_k$
$\rho_{ik}=-½$

$C_{ik}=0$
$\rho_{ik}=0$

$C_{ik}=+½\sigma_i\sigma_k$
$\rho_{ik}=+½$

$C_{ik}=\sigma_i\sigma_k$
$\rho_{ik}=+1$

Materials in this slide adapted from Dr. Ricardo Gutierrez, Texas A&M, USA

Gaussian Processes

Definition:

- A Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian Distribution.
- A GP is completely specified by its mean function and the covariance function.

$$m(\boldsymbol{x}) = E[f(\boldsymbol{x})]$$

$$Cov(\boldsymbol{x_p}, \boldsymbol{x_q}) = \mathrm{E}\left[\left(f(\boldsymbol{x}_p) - m(\boldsymbol{x}_p)\right)\left(f(\boldsymbol{x}_q) - m(\boldsymbol{x}_q)\right)\right]$$

GP for Spatial Interpolation

- GP is a collection of random variables continuously distributed over the spatial domain
- GP can be considered as a random function that interpolates over the domain.

Random multivariate images

- **Additional hypothesis**
- **Time variaion**
  - We do not consider the time variation
  - Time can be considered as an additional dimensión with the same formalism
- **Stationarity.**
  - Correlation among random variables depend on just the distance, but not the particular point in the map.

$$Cov(\boldsymbol{h}) = \mathrm{E}\left[\left(f\left(\boldsymbol{x}_p\right) - m\right)\left(f\left(\boldsymbol{x}_p + \boldsymbol{h}\right) - m\right)\right]$$

  - In the basic formulation the mean is assumed to be constant over the Surface. If we suspect it is not, usually a smooth baseline model is estimated and substracted from data. Then the same analysis is applied.
- **Ergodicity**
  - We asume that averaging accross samples in space is the same as averaging across realizations.

Empirical Semivariance

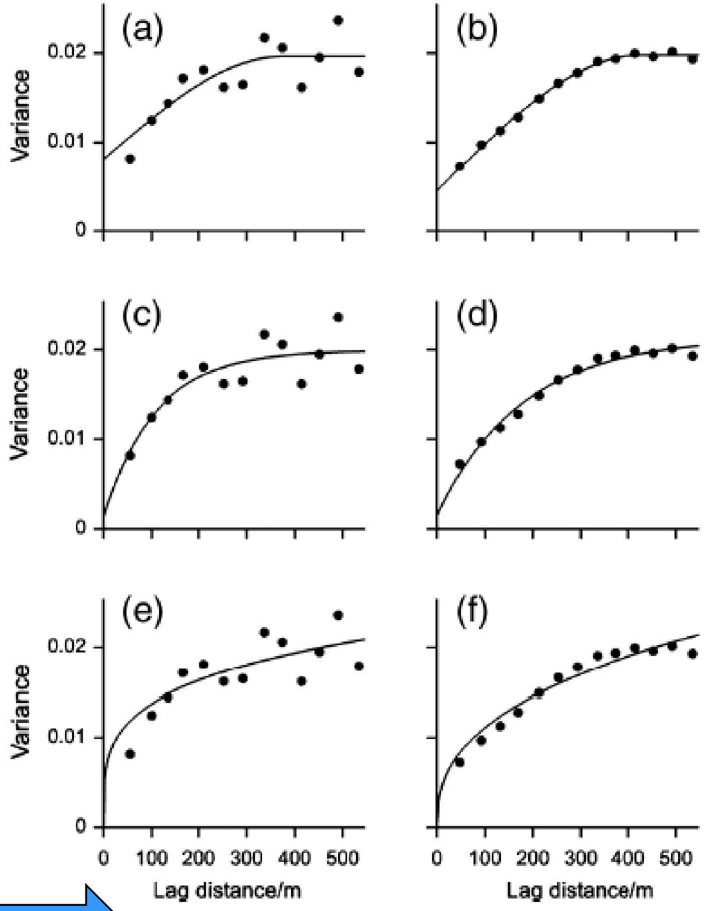
- **Instead of using the covariance, the semivariance is often used to characterize the space variability. Both of them are related.**

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\mathrm{Var}[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})]$$

$$\gamma(\boldsymbol{h}) = \mathrm{Cov}(\mathbf{0})\text{-}\mathrm{Cov}(\mathbf{h})$$

- **The representation of the semivariance agains distance is called the variogram.**

Experimental variograms of log10[K] in the soil in a farm.
Left column 87 measurements
Right column 434 measurements
Minimum distance 40 m.
Variograms have been fitted to diverse models.(Oliver, Webster, 2013)



Fitting the variogram

- **In order to properly model the Gaussian Process (Also named Kriging) we need to know the variogram for any distance.**
- **In order to do so, analytical functions are fitted to the empirical variogram.**
- **The simpler (and often used) models are spherical.**
- **The most common model is:**

$$\gamma(h) = c_0 + \mathrm{c}\left\{\frac{3h}{2r} - \frac{1}{2}\left(\frac{h}{r}\right)^3\right\} \qquad \text{for h<r}$$

$$c_0 + \mathrm{c} \;\; \text{for h} \geq r$$

$$0 \text{ for h=0}$$

Where h is the distance, c_0 is called the nugget variance it respresents the variance measurement, c_0 +c is the process variance.

Alternative models for the variogram

- **Exponential:**

$$\gamma(h) = c_0 + c\left\{1 - exp\left(-\frac{h}{a}\right)\right\}$$

$$\gamma(h) = 0 \qquad h = 0$$

- **Power**

$$\gamma(h) = c_0 + bh^{\eta} \qquad \text{with } 0<\eta < 2$$

$$\gamma(h) = 0 \qquad h = 0$$

- **Gaussian:**

$$\gamma(h) = c_0 + c\left\{1 - exp\left(-\frac{h^2}{a^2}\right)\right\}$$

$$\gamma(h) = 0 \qquad h = 0$$

- **Gaussian variograms are not recommended since they lead to over smooth maps**

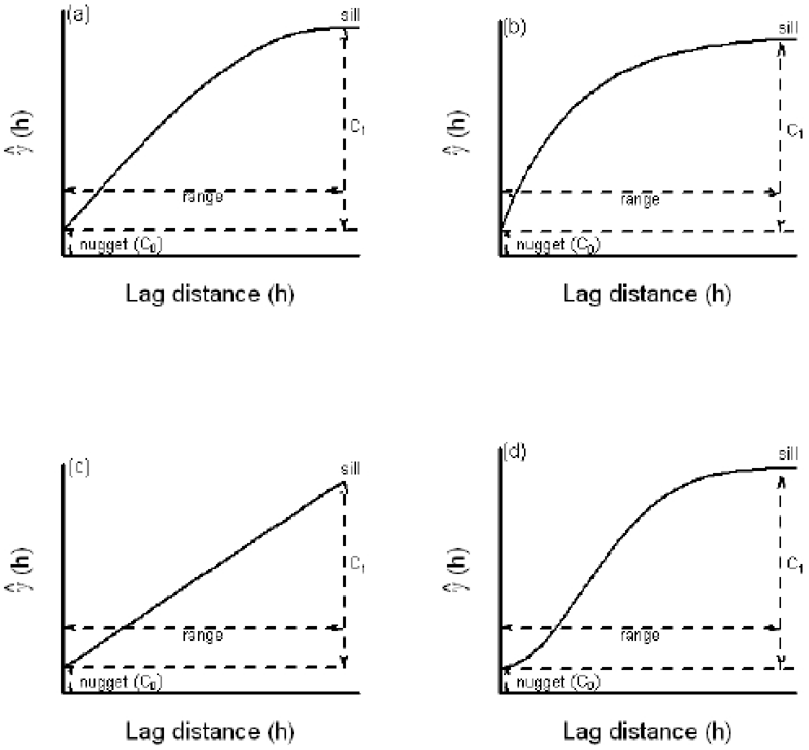
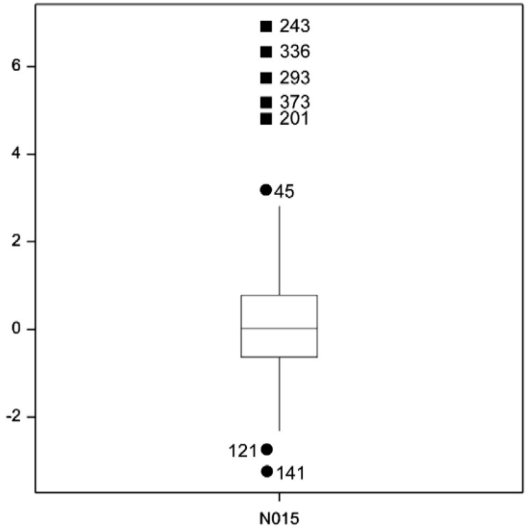
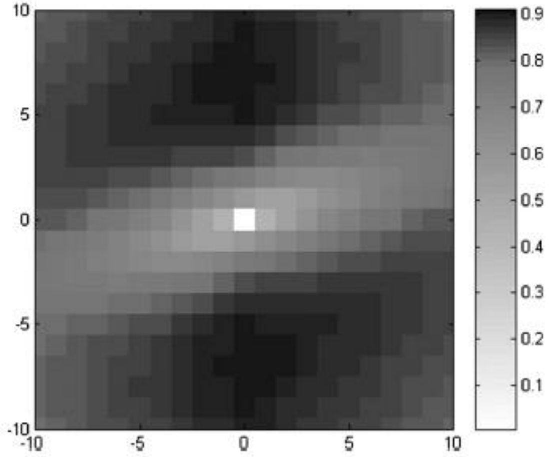


Figure 2.2. Examples of four commonly used variogram models: (a) spherical; (b) exponential; (c) linear; and (d) Gaussian.

Issues on the estimatio of the variogram

- **Sample size:**
  - A the amount of samples over the domain impact the quality of the variogram. Few samples will provide a large variance in the variogram.
- **Distance between samples and domain size.**
  - If the domain size is too small, we will not see the asimptotic variation of the correlation
  - If the sample distance is too large maybe we cannot grap the most important change in the variogram.
- **Distribution of samples.**
  - Its important that samples are gaussian distributed. If not proper non-linear transformation are recommended to improve variogram computation. Typical transformation are logarithms or power functions. A number of ways to determine if data is gaussian exist but the easiest thing is too do a boxplot. This also helps to identify outliers.
- **Anisotropy**
  - Sometimes there are preferential directions. To identify this the variogram can be estimated in different directions of h. If important diferences arise an anisotropic model may improve map.





Prediction equations for the Gaussian Process.

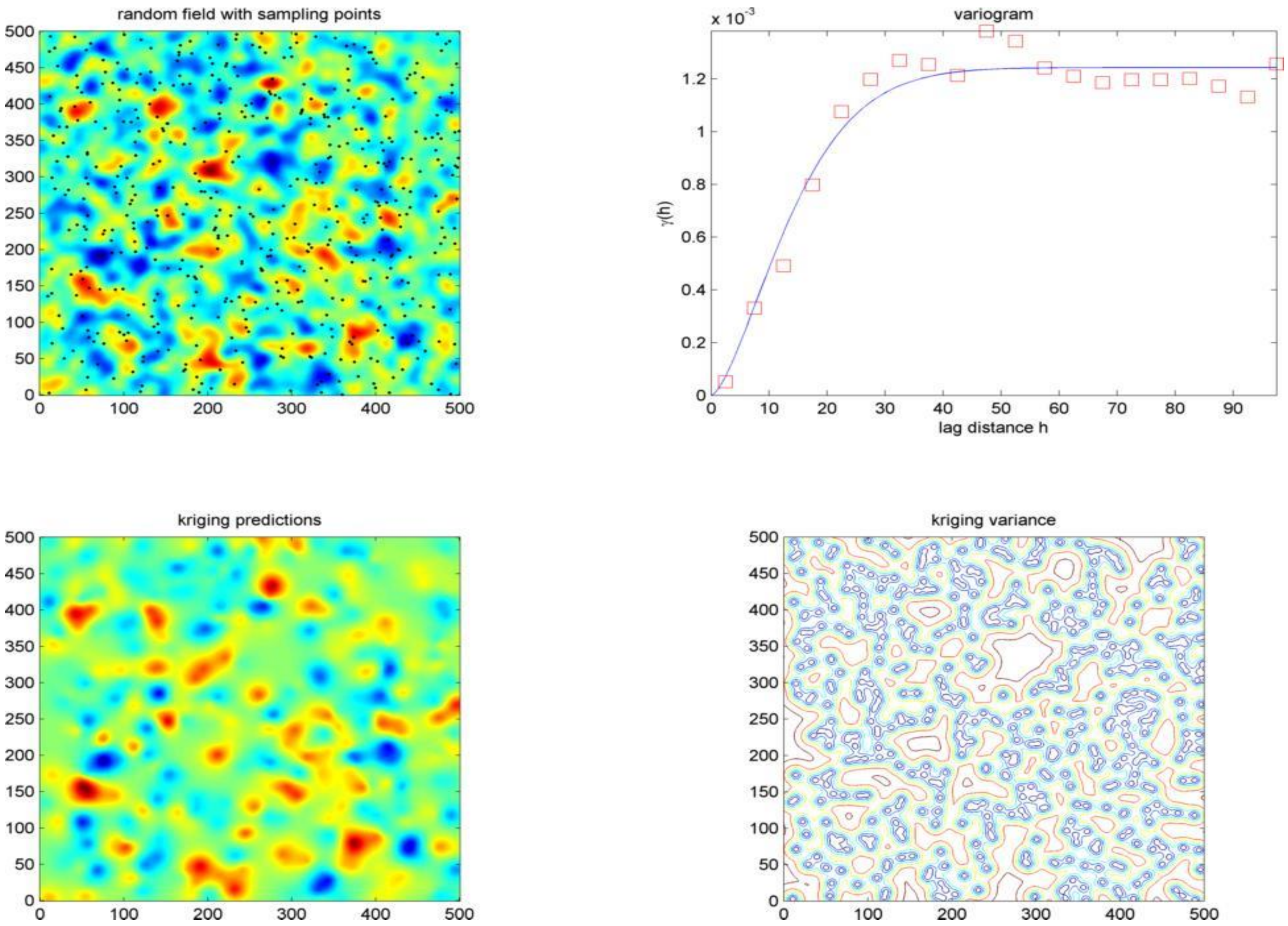
- **K(X,X) is the covariance matrix of the sample set data (it can be extracted from the variogram).**
- **σ is the measurement noise.**
- **K* is the vector of covariances between the prediction point and all the others in the training set.**
- **Y is the vector of prediction points in the training set.**

$$\widehat{f_*} = k_*^T(K + \sigma_n^2 I)^{-1} Y$$

$$\text{Var}(f_*) = k(x_*, x_*) - k_*^T(K + \sigma_n^2 I)^{-1} k_*$$

- **Note that the prediction is a weighted combination of Y values. The dependence on the point is hiden in k*.**
- **The technique additionally provide the variance of the point.**
- **The technique uses the empirical variogram to account for the shape of the existing**

Example of results with GP / Kriging



http://www.mathworks.com/matlabcentral/fileexchange/29025-ordinary-kriging

Evaluation of the results

- The evaluation of interpolation quality is often done by the same methods used to evaluate regression models.
- The main technique is Cross-Validation: either k-fold and Leave-one-out.
- Typical figures of merit are:

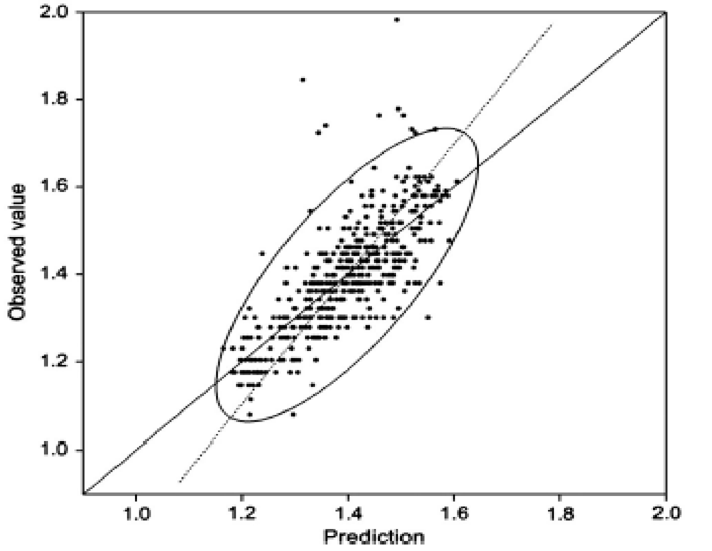


Fig. 11. Scatter diagram of observed values of $\log_{10}K^{+}$ at Broom's Barn plotted against values predicted by ordinary punctual kriging during cross-validation. The solid line (approximately 1:1) is the regression of observed on predicted values, the dashed line is the principal axis. The variances are 0.01800 and 0.00990 respectively.
From Webster and Oliver (2007).

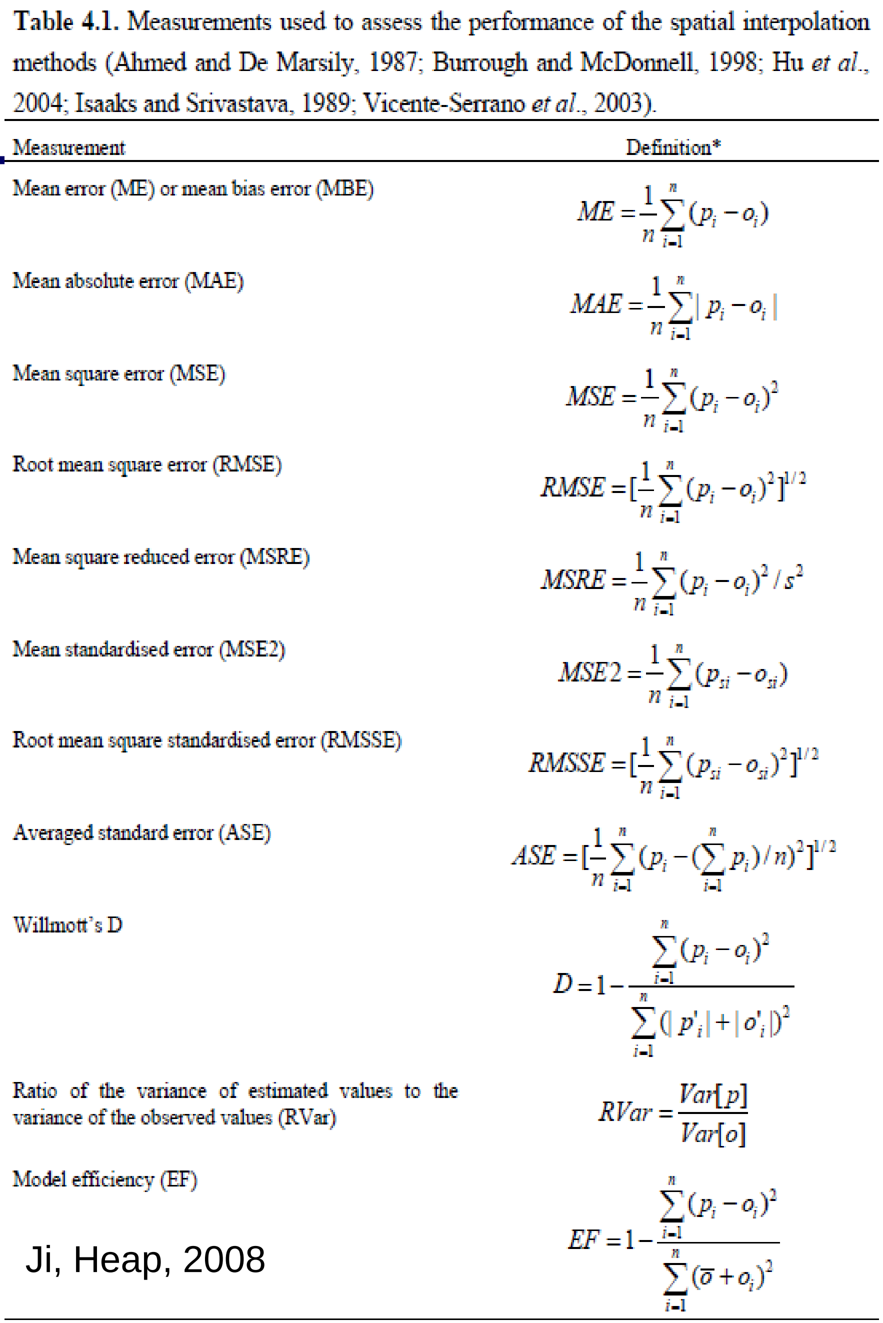


Table 4.1. Measurements used to assess the performance of the spatial interpolation methods (Ahmed and De Marsily, 1987; Burrough and McDonnell, 1998; Hu *et al.*, 2004; Isaaks and Srivastava, 1989; Vicente-Serrano *et al.*, 2003).

Measurement	Definition*
Mean error (ME) or mean bias error (MBE)	$ME = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)$
Mean absolute error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert p_i - o_i \rvert$
Mean square error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$
Root mean square error (RMSE)	$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2]^{1/2}$
Mean square reduced error (MSRE)	$MSRE = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2 / s^2$
Mean standardised error (MSE2)	$MSE2 = \frac{1}{n}\sum_{i=1}^{n}(p_{si} - o_{si})$
Root mean square standardised error (RMSSE)	$RMSSE = [\frac{1}{n}\sum_{i=1}^{n}(p_{si} - o_{si})^2]^{1/2}$
Averaged standard error (ASE)	$ASE = [\frac{1}{n}\sum_{i=1}^{n}(p_i - (\sum_{i=1}^{n} p_i)/n)^2]^{1/2}$
Willmott's D	$D = 1 - \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(\lvert p'_i \rvert + \lvert o'_i \rvert)^2}$
Ratio of the variance of estimated values to the variance of the observed values (RVar)	$RVar = \frac{Var[p]}{Var[o]}$
Model efficiency (EF)	$EF = 1 - \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(\overline{o} + o_i)^2}$

Ji, Heap, 2008

Tools for Spatial Interpolation

- **Many tools are available in the public domain:**
- **GSLIB**
  - GSLIB is a collection of routines developed in Stanford University during the 90s.
  - The source code is in FORTRAN 90
  - Routines description is available as a book
  - http://www.gslib.com
- **GSTAT in R**
  - http://cran.r-project.org/web/packages/gstat/index.html.
- **GSTAT in MATLAB**
  - http://mgstat.sourceforge.net/
- **PYSAL in Python**
  - https://pypi.python.org/pypi/PySAL
- Some geographic information systems (GIS) have built in options for spatial interpolation. For instance the 'Spatial Analyst' in ArcGIS ver. 9.2.
- General remark:
  - Obtaining interpolation maps with current software can be just some clicks away but the use of the algorithms as black boxes can lead to bad estimations particularly in sparsely sampled domains.

Summary

- There is a growing trend to have more distributed environmental data due to low cost sensor nodes.
- In many occasions the final goal is to have a map over a whole domain of the variable of interest.
- Many methods exists coming from different disciplines and with different assumptions, but the field is known today as *Spatial Statistics.*
- While many methods may work, the recommended method today is known as Kriging or Gaussian Processes.
- Ready available tools are available in several programming languages and softwares
- The user needs to know basic concepts about the algorithms to understand what he/she is doing.

References

- Jin Li, Andrew D. Heap, “A Review of Spatial Interpolation Methods for Environmental Scientists”, Australian Gov GeoCat# 68229. (2008).
- M. A. Oliver, R. Webster, “A tutorial guide to Geostatistics: Computing and Modeling Variograms and Kriging”, Catena, 113 (2014) 56-69.
- R.S. Bivand, E.J. Pebesma, V. Gómez-Rubio, “ Applied Spatial Data Analysis with R”, Springer, 2008.
- C. E. Rasmussen, C. K.I. Williams, “Gaussian Processes for Machine Learning”, MIT Press, 2006.