Data Processing

University "UKSHIN HOTI" Prizren

Co-funded by the Horizon 2020 programme of the European Union

DT-ICT-01-2019
Smart Anything Everywhere Area 2

www.smart4all-project.eu
Grant Agreement: 872614

Content:

- Introduction
- Statistical approach to data fusion
  - Geostatistical tools
  - Kriging
  - Cross-validation
  - Multivariate methods
  - Geostatistical approaches to data fusion

Introduction

- Land managers need efficient tools to analyze and manage huge datasets.
- Efficient techniques for processing and summarizing data will be crucial for effective management.
- Spatial data often turn out to be 'incompatible' owing to their heterogeneities in terms of nature (continuous or categorical), quality (soft or hard data), and spatial and temporal scales.
- Complex spatial dependence and interdependence structures among spatial variables contribute to making data fusion more difficult.

Statistical approach to data fusion

- The support of spatial data is the physical volume over which the value of a variable is measured or computed.
- In geostatistics, the concept of a regularized variable is one of the key ideas and is strongly related to support. It represents the average value over a volume *v* of a variable *Z* defined on a point volume:

$$Z(v) = \frac{1}{|v|}\int_v Z(\mathbf{x})d\mathbf{x}$$

- where $|v|$ is called the support of $Z(v)$ and x is the vector of two (2D) or three (3D) coordinates (x_1, x_2, x_3).

Statistical approach to data fusion

- Many practical applications in Precision Agriculture use data measured on small samples to estimate the variable of interest over a much bigger unit (field or farm).
- Data can be associated with lines, surfaces, or volumes of any shape.
- Developing a proper methodology for upscaling and downscaling the data is crucial for accurate inference.

The ‘change of support’

- Spatial fields generally have a spectrum of variability including different spatial scales: from long range to microscopic scale.
- The support effect is crucial when estimating the conditional probability that the average over a specified area is below a critical value.
- In agriculture, such a probability could be used to decide if nutrients/water need(s) to be added to the soil.

The ‘change of support’

- Change of support problem (COSP) involves two inferential issues.
- The shape of a variable averaged over spatial units is different from the one of the original variable.
- Aggregation then tends to reduce heterogeneity among the units, although it is more complicated by spatial autocorrelation.
- Inferences obtained at a given scale cannot be transferred to another scale.

Different solutions to COSP

- Solving COSP is intended as the ability to make spatial predictions on a given target (target support) by using data associated with a set of supports (source supports).
- An approach should satisfy the following main requirements:
  - explicitly accounting for the different supports involved;
  - being able for upscaling from points to volumes;
  - accounting for the uncertainties of the source data;
  - integrating covariates of any type to improve predictions;
  - preserving consistency in predictions across the scales;
  - being easily implemented within a GIS to perform calculations involving point-to-point.

GIS operations

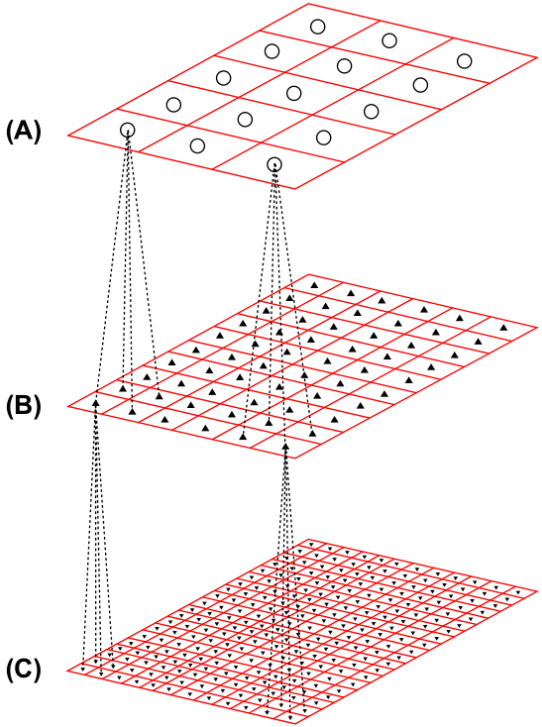
- Data in GIS applications are in either vector form or raster form.
- Typical GIS operations are union, intersection, and dissolve (on vectorial data) or zonal averaging and pixel-by-pixel computations on raster data.
- All these operations can be implicitly considered as fusion operations.

Geographically weighted regression

- Geographically weighted regression (GWR) differs from standard linear regression for two aspects.
- It assumes that nearby observations are more similar than those far apart.
- GWR is a very common but univariate method of spatial interpolation, although it could be extended to multiple sources.

Multiscale spatial tree models

- Multiscale tree models have several advantages because they can work with large datasets (Big Data).
- They are flexible to be used in a wide range of applications.
- The main disadvantage consists in the fact that they do not treat explicitly COSP associated with changes in resolution.



Bayesian hierarchical models

- The most classical techniques of fusing information are based on probability theory associated with Bayesian decision theory.
- A simple rule for fusing data at the prediction location is assuming mutual independence for information at other locations.

Geostatistical tools

- Geostatistics aims at providing quantitative descriptions of variables distributed in space or in time and space.
- Examples of geostatistical applications can be found in environmental and soil sciences, meteorology, hydrology, ecology, remote sensing, fisheries, public health, and also in PA.

Random function and regionalized variables

- Classic statistics assume that the expected value of a given soil property z, at a point x within the sampling area, is given by

$$z(x) = \mu + \varepsilon(x)$$

- where μ is the mean of the population and
- $\varepsilon(x)$ is a random variable.
- Classic statistical procedures assume, therefore, that the variation within an area is randomly distributed.
- Regionalized variables established a better way to represent the reality, introducing randomness in terms of fluctuation around a fixed surface.

The variogram

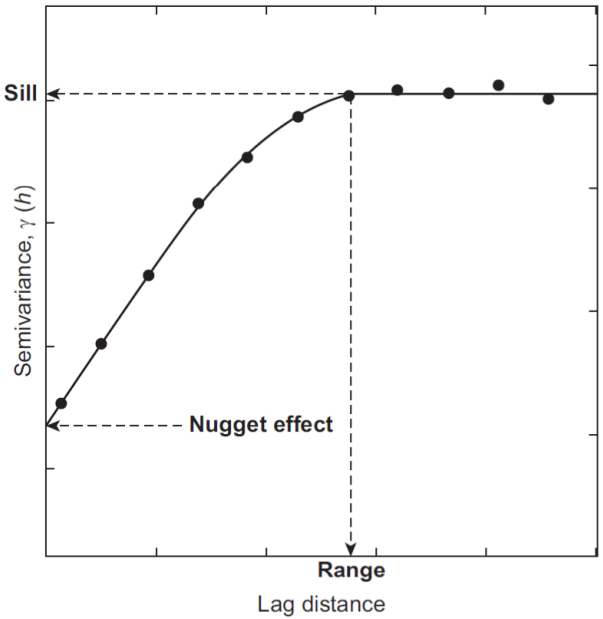
- The semivariance, $\gamma(\mathbf{h})$, represents the spatially dependent component of the random function Z.

$$\gamma(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{\alpha=1}^{n(\mathbf{h})} [z(\mathbf{x}_\alpha) - z(\mathbf{x}_\alpha + \mathbf{h})]^2 \quad \alpha = 1, \ldots, n(\mathbf{h})$$

- where $n(\mathbf{h})$ represents the number of pairs separated by the same lag distance.
- In the case of second-order stationarity, it can be expressed in terms of spatial covariance, $C(\mathbf{h})$, and spatial variance, $C(0)$ of a regionalized variable.

The variogram

- For each selected direction, the semivariance is generally represented by a graph of $\gamma(\boldsymbol{h})$ as a function of $\boldsymbol{h}$.
- Variograms must be conditionally negative definite functions.
- A high nugget value indicates a lack of spatial correlation.
- Variograms must be conditionally negative definite functions.



Kriging

- Kriging is a linear technique that allows us to get unbiased estimates of a regionalized variable in unsampled points.
- The kriging estimator is said BLUE (best linear unbiased estimator) for its properties.
- Kriging calculates an error term (estimation variance) for each estimated value, thus providing a measure of the reliability of interpolation.

Cross-validation

- Cross-validation is a procedure to check compatibility between the data and the model.
- It involves eliminating one data point and estimating its value using the remaining data with kriging.
- Each estimate is compared with the measured value by calculating the experimental error, i.e. the difference between estimate and measurement.

Multivariate methods

- In many cases, spatial studies consider two or more variables.
- A cross-variogram describes as the variable i is spatially related to the variable j.
- The higher the correlation between the two variables, the more similar the two direct (or auto) variograms are.
- Linear Model of Coregionalization (LMC) assumes that all variables are a linear combination of the same basic structures, each one corresponding to a given spatial scale related to a specific process.

Cokriging

- Cokriging is the multivariate extension of kriging formalism. It allows dealing simultaneously with two or more variables defined over the same domain.
- Like kriging, cokriging is quite flexible and applicable to a wide variety of problems.
  - It may require the inversion of large matrices to solve the linear system of equations, which makes it computationally prohibitive with large datasets (Big Data).
- An alternative is multicollocated cokrigging, which utilizes for prediction only the exhaustive variable(s).

Kriging with external drift

- When the assumption of spatial stationarity does not hold for the variable of interest, alternative solutions have to be used.
- The basic hypothesis of kriging with external drift is that the expectation of the variable can be modeled as the sum of polynomials and linear functions of secondary variables.

Geostatistical approaches to data fusion and COSP

- The key idea of kriging is that any block variogram (covariance) of z(B) can be determined from the underlying point process z(s).
- The method can be used for upscaling (aggregation), downscaling, and side scaling.
- COSP is based on the knowledge (from observations) or calculation of the point covariance (variogram) function.
- Cokriging equations contain the direct and cross-covariance functions of each variable.

Application of geostatistical data fusion in proximal sensing

- Satellites collect data over discrete areal regions called footprints, which represent aggregated views of underlying continuous processes. A satellite can survey only a limited portion of the space-time domain.
- Many methods have been proposed for image fusion, including the intensity-hue-saturation (HIS) method.
- Cokriging explicitly takes into account the pixel size (support) of each image.
  - A typical application of cokriging is processing coregistered images with different spatial resolutions in the different spectral bands.
  - An example is using Sentinel-2 satellite sensor images to increase the spatial coarse resolution (60 m) of the bands 1, 9, and 10.

Application of geostatistical data fusion in proximal sensing

- The problem of predicting a fine spatial resolution image with cokriging can be solved by fusing images with different spatial resolutions in different spectral bands.
- The method implies numerical deconvolutions for estimating the covariances and cross-covariances with point support.

Application of geostatistical data fusion in proximal sensing

- PA is based on the assumption that optimum benefits on profitability and environmental protection depend on the level of agreement between agricultural practices and local conditions.
- It is quite critical for PA to assess spatial and temporal variation of soil/plant accurately and locally at a very fine scale.
- New approaches are needed to analyze massive datasets more efficiently.
- Multicollocated cokriging can make such a data fusion possible.

Application of geostatistical data fusion in proximal sensing

- When various complementary sensors are available (Fig), it is expected that sensor data fusion or data integration may perform inferences potentially more accurately than the ones achieved by a single sensor.

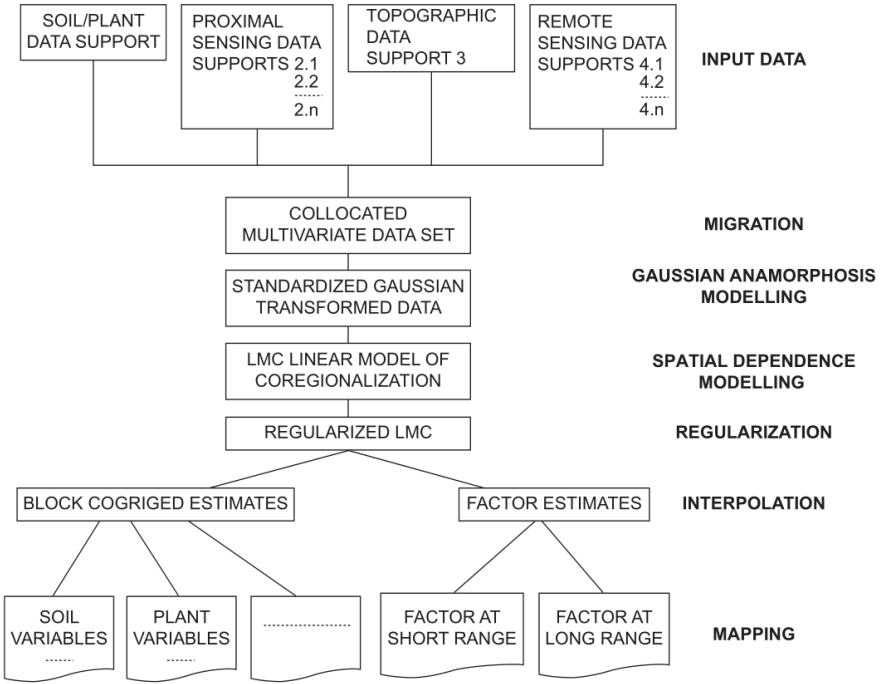
Multisensor platform at CREA-AA, Italy

Data fusion geostatistical approach

- The main procedures required by a geostatistical approach to fuse multitype datasets:
  - *Sample data migration*
  - *Gaussian anamorphosis modeling*
  - *LMC fitting*
  - *LMC regularizing on block support*
  - *Block cokriging performing*
  - *Factorial block cokriging performing*

Data fusion geostatistical approach

Flowchart of the proposed geostatistical data fusion approach.

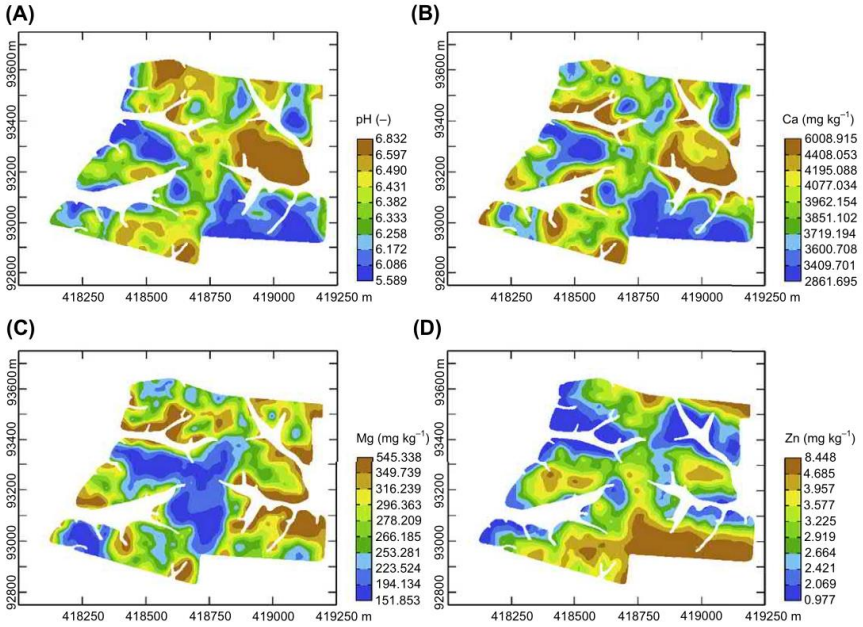


Case study

- Case study aims to investigate the scale-dependent correlation structure of a multivariate input dataset, including some soil variables, fine-scale terrain data, and geophysical measurements of soil bulk electrical conductivity.
- Main objectives are to provide a thematic map of soil attributes and to determine a few spatial scale-dependent indices for the delineation of homogeneous within-field areas.

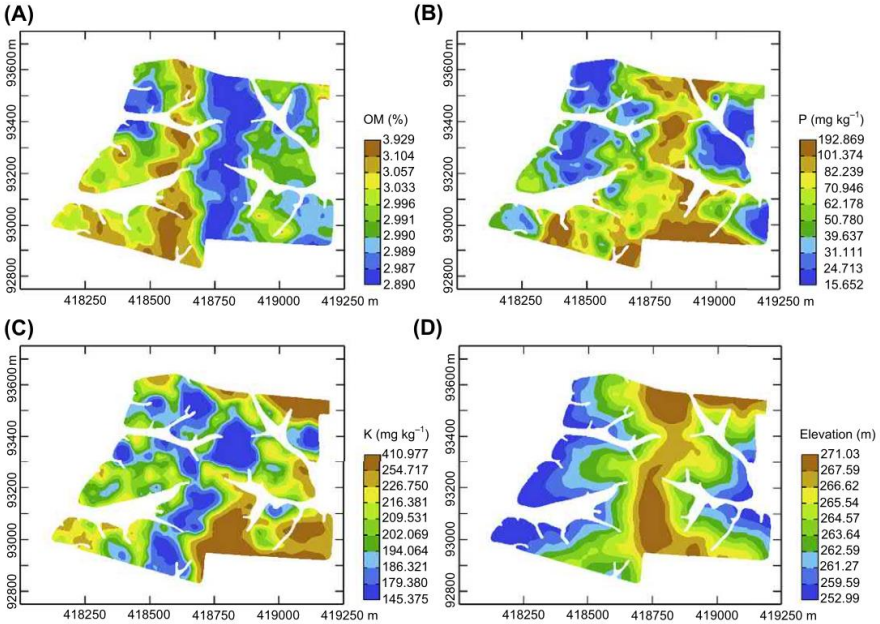
Case study

Block cokriging maps of pH (A), Ca (B), Mg (C) and Zn (D).



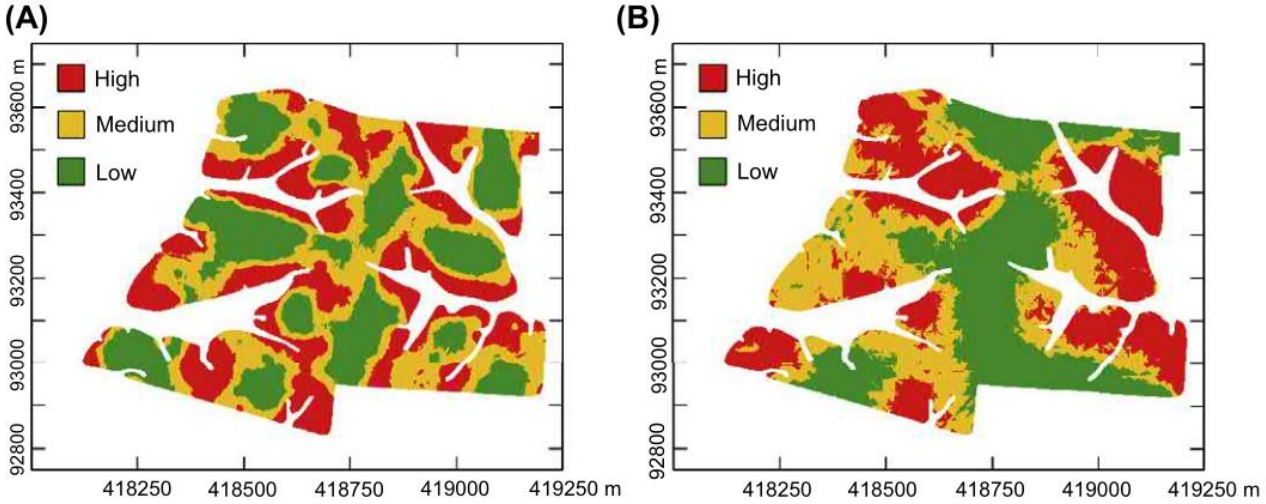
Case study

Block cokriging maps of (A) organic matter (OM), (B) phosphorous (P), (C) potassium (K), and (D) elevation.



Case study

Regionalized factors at (A) short-range (140 m) and (B) long-range (500 m).



Session Q&A