1

Use of Descriptive Statistical Indicators for Aggregating Environmental Data in Multi-Scale European Databases

Panos Panagos, Yusuf Yigini and Luca Montanarella
Joint Research Centre of the European Commission,
Institute for Environment and Sustainability,
Italy

1. Introduction

1.1 Policy context

There is a strong need for accurate and spatially referenced information regarding policy making and model linkage. This need has been expressed by land users, and policy and decision makers in order to estimate spatially and locally the impacts of European policy (like the Common Agricultural Policy) and/or global changes on economic agents and consequently on natural resources (Cantelaube et al., 2012).

The proposal for a framework Directive (COM (2006) 232) (EC, 2006) sets out common principles for protecting soils across the EU. Within this common framework, the EU Member States will be in a position to decide how best to protect soil and how use it in a sustainable way on their own territory. In this policy document, European Commission identifies 8 soil threats: soil erosion, soil organic carbon decline, salinisation, landslides, soil compaction, biodiversity and soil contamination. The policy document explains why EU action is needed to ensure a high level of soil protection, and what kind of measures must be taken. As the soil threats have been described in the proposed Soil Thematic Strategy for Soil Protection (COM (2006) 231), there is a need to address them and relative issues at various scales; from local/province scale, to regional/national scale, and at the end to continental/global scale. The modeling platform should be constructed in such a way that knowledge and information can be passed along the spatial scales causing the minimum loss of information. Particular interest will be given to outputs from the aggregation model such as organic carbon decline, soil erosion and soil.

The INSPIRE Directive (INSPIRE, 2007) aims at making relevant geographic information available and structurally interoperable for the purpose of formulation, implementation, monitoring and evaluation of Community policy-making related to the environment. To that end, data specifications for various themes are to be developed. The Soil theme is listed in Annex III of the INSPIRE Directive.

Soil organic data are requested for models relating to climate change. The role of soil in this debate, in particular peat, as a store of carbon and its role in managing terrestrial fluxes of

carbon dioxide (CO2), has become prominent. Soil contains about twice as much organic carbon as aboveground vegetation. Soil organic carbon stocks in the EU-27 are estimated to be around 75 billion tonnes of carbon (Jones et al., 2005).

Soil data and information are highly relevant for the development, implementation and assessment of a number of EU policy areas: agriculture, soil protection, bio-energy, water protection, nature protection, development policy, health and sustainable development. All those policy areas request soil data in various scales depending on the application.

Regarding research purposes, according to the data logs in European Soil Data Centre (Panagos et al., 2012), the users deploy ESDAC data mainly (but not exclusively) for modeling purposes (35%). Most of the modelling exercises request the input data to be transferred in a specific scale in order to fit the modeling purposes. Most of the modeling is performed in small scales covering few square kilometres; however, during the last years the modeling exercises performed in national or European level is increasing due to high demand for environmental indicators performance.

1.2 Multi-scale European Soil Information System (MEUSIS)

Implementation of the INSPIRE directive should emerge the development of a Multi-scale European Soil Information System (MEUSIS), from the data producer up to the final user, responding to the various needs at different scales. In order to achieve this, a common standard for the collection of harmonized soil information will have to be implemented. As a response to this requirement, MEUSIS is proposed as a harmonized hierarchical Grid (Raster) data system which constitutes an ideal framework for the building of a nested system of soil data. This reference grid is based on implementing rules facilitating data interoperability.

The final result of these developments should be the operation of a harmonized soil information system for Europe streamlining the flow of information from the data producer at a local scale to the data users at the more general Regional, National, European and Global scales. Such a system should facilitate the derivation of data needed for the regular reporting about the state of European soils by European Commission authorities.

However, soil geography, soil qualities and soil degradation processes are highly variable in Europe. Soil data sets from different countries have been often created using different nomenclatures and measuring techniques, which is at the origin of current difficulties with comparability of soil data. The availability of soil data is also extremely variable in Europe. Individual Member States have taken different initiatives on soil protection aimed at those soil degradation processes they considered to be a priority.

Traditionally, the European Soil Database has been distributed in vector format. More recently, interest was expressed for deriving a raster version of this database. In the specific case of MEUSIS, the advantages of the raster approach are listed below:

- Easy to identify the data per location. Each cell has an ID and its geographic location is determined by its position in the matrix cell.
- It is fairly easy to store data and to perform data analysis.

- It is easy to integrate data from different data sources or different data types. As a result soil data could be processed by other environmental indicators and can be imported in data models such as climate change ones.
- The pixel approach would make it easier for data to be updated.
- The structure is suitable to perform upscaling (bottom-up) from local to regional, national and European level.

The main disadvantage of the raster approach is that this technique is less precise in representing the real world, which means that it is not suitable for representing soil coverage complexity and it might not be always easy to persuade the general public about the potential usability of this technique. In Figure 1 portray an example on how pixel cells of 1km² size may be represented in a higher resolution grid or raster of 10 km².

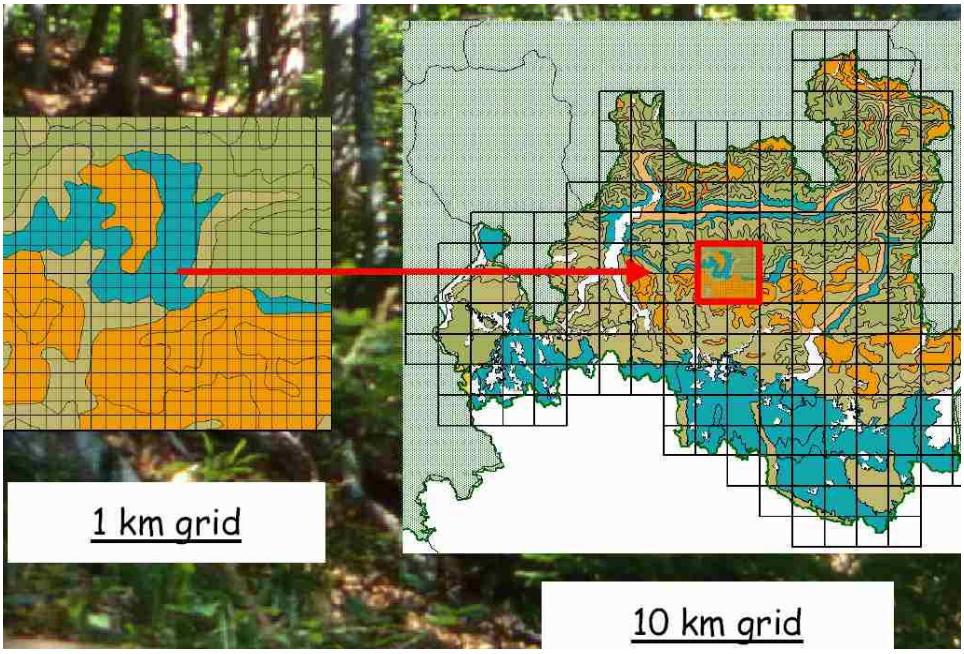


Fig. 1. Grid Example in 2 different resolutions

2. Upscaling

Upscaling of environmental indicators applied in regional analyses is sensitive to scale issues of the input Data (Bechini et al., 2011). Environmental assessments are frequently carried out with indicators (Viglizzo et al., 2006) and simulation models (Saffih- Hdadi and Mary, 2008). The environmental indicators have an increasing importance and are easily understandable by the general public. Those quantitative expressions measure the condition of a particular environmental attribute in relation to thresholds set by scientific community. However, decision makers use the environmental indicators to communicate with the general public.

When dealing with areas of different sizes and with information available at different scales, policy makers and decision makers need to either upscale their evaluations and simulations from small to large scale or downscale from large to small scale (Stein et al., 2001). Environmental indicators are dependent upon data availability and also upon the scale for which policy statements are required. As these may not match, changes in scales may be necessary. Moreover, change is scale may requested in research and modeling where the indicator is used as input parameter in a model. It has been recognised that the quality of indicators relies on the scale which they represent. The quality of the state of the environment at a local scale, for example, requires different information compared to the state of the environment at national scale.

From the one hand, ecologists criticize upscaling approaches insisting that it ecological knowledge is difficult to scale up (Ehleringer and Field, 1993). They support that environmental systems are organized hierarchically with multiple processes taking place across scales. When moving from a finer scale to a coarser one in this nested hierarchy, new processes may be encountered which is difficult to be translated in research results. The environmental systems are not non linear ones and no scaling rules can be imposed to express such a behaviour. Environmental systems are spatially heterogeneous due to spatial variations in climatic and soil conditions. As you can see from the references, this was mostly the trend in the 80's-90's while in the recent years there are many applications of upscaling in many environmental fields.

Scale for environmental indicators has barely been addressed in the literature. Scale issues are considered to be of importance (Bierkens et al., 2000) and advantages have been reported in hydrology (Feddes, 1995) and soil science (Hoosbeek and Bouma, 1998; McBratney, 1998). Upscaling is the process of aggregating information collected at a fine scale towards a coarser scale (Van Bodegom et al., 2002). Downscaling is the process of detailing information collected at a coarse scale towards a finer scale.

Scale is defined as the spatial resolution of the data. Scales, defined in terms of resolution and procedures, are presented to translate data from one scale to another: upscaling to change from high resolution data towards a low resolution, and downscaling for the inverse process. Environmental assessments at a small scale commonly rely on measured input, whereas assessments at a large scale are mainly based on estimated inputs that cannot be measured or outputs of modeling exercises.

Policy makers request to know also the uncertainty of environmental assessments in order to better interpret the results and proceed with the most suitable decision. The quantification of uncertainty implies the confidence level of indicators which can be measured with statistical measurement such as standard deviation.

Upscaling in complexity means that data quality degrades with decreasing complexity, because information is generalised and uncertainty increases. In literature, upscaling is defined as the process that replaces a heterogeneous domain with a homogeneous one in such a manner that both domains produce the same response under some upscaled boundary conditions (Rubin, 1993). The difficulty in upscaling stems from the inherent spatial variability of soil properties and their often nonlinear dependence on state variables. In 2004, Harter and Hopmans have distinguished four different scales: pore scale, local (macroscopic), field and regional (watershed). In this study the upscaled processes are performed between 3 scales: local, regional and national.

The scaling methods are applied before the geostatistical analysis in order to avoid dealing with multiple, spatially variable but correlated physical quantities. Environmental modelling requires the input spatial data to be in the same scale and upscaling/downscaling processes assist in transferring the input data in the requested scale. Geostatistics is used to make predictions of attributes at un-sampled locations from sparse auxiliary data. Upscaling is also used in disciplines or applications where there may be too much data which need to reduced to manageable proportions.

Based on King's approach for explicit upscaling in space (King, 1991), we will try to integrate the heterogeneity that accompanies the change in model extent by averaging across heterogeneity in the soil organic carbon data and calculating mean values for the model's arguments.

3. Material and methods

3.1 Indicators – Organic carbon

An environmental indicator is defined as a measure to evaluate or describe an environmental system. The indicator should be measurable and the threshold values attached to it would facilitate its presentation to the public. The indicators require to a scientific background and a sound method of evaluation (Gaunt et al., 1997). One of the main characteristics for the definition of an environmental indicator is the application in space and time. In this context, the indicator can be aggregated to a more coarse scale in order to serve decision making. Here, comes the contribution of statistics in comparing the indicators by using specific figures such as mean, median, mode, standard deviation, sample variance, quartile, ranges, etc.

Soil research and policy makers in the soil field needs statistics to support and confirm the impressions and interpretations of investigations in the field. The use of mathematics and statistics becomes more and more popular among soil scientists. The terms such as geostatistics become popular in the soil science community while new software tools facilitate such data processing with the help of more powerful computers.

However, Minasny and McBratney argued that better prediction of soil properties can be achieved more with gathering higher quality data than using sophisticated geostatistical methods and tools. However, it should be underlined the high cost and the time consuming for laboratory analysis of field data; that is why research in developing methods for the creation of soil maps from sparse soil data is becoming increasingly important. In the last 20 years, the development of prediction methods using cheap auxiliary data to spatially extend sparse and expensive soil information has become a focus of research in digital soil mapping (Minasny and McBratney, 2007). Examples of secondary information, named covariates, include remote sensing images, elevation data, land cover and crop yield data.

In order to describe the upscaling methodology, a data field such as the Organic Carbon (OC) content in the surface horizon 0-30 cm of the Slovakia Soil Database will be used. The Organic Carbon is a quantitative attribute measured as tones per hectare according to the following equation:

$$OC(t/ha) = Cox * BD * d$$

Where,
Cox (%) is the average content of organic carbon for topsoil/subsoil,
BD (g/cm^3) is the average soil bulk density for topsoil/subsoil,
d (cm) is the volume of topsoil/subsoil

Soil organic carbon is an important soil component as it influences soil structure and aggregation, soil moisture conditions, soil nutrient status and soil biota, and hence influences ecosystem functioning (Lal, 2004).

3.2 Changes in scale

Spatial scale refers to the representativeness of the singe measurements (or observations) for larger mapping units. The level of variation is different depending on the scale; few measurements at a coarse scale in a large area have a different variation from few measurements in a fine scale or many measurements in a large scale. Upscaling is the process of changing scale from fine to coarser one and it is performed with procedures such as averaging or block kriging. Use of confidence levels and ranges appears useful in upscaling. The use of GIS advanced systems is useful to visualise the affects of upscaled result and contribute better t communication with public and decision makers.

3.3 Aggregation technique and cell factor

Scale factors in general are defined as conversion factors that relate the characteristics of one system to the corresponding characteristics of another system (Tillotson and Nielsen, 1984). Aggregating functions in the upscaling methodology and spatial data process will be done using ArcGIS software. As a GIS technique, spatial join is proposed since spatial data from one layer can be aggregated and added to objects of the other layer, which is often referred to as the destination layer. Aggregation is accomplished via a cell fit criterion since many data cells from one source layer would fit in one cell in the destination layer. The modeller must decide how existing attributes will be summarized during aggregation (e.g., averages, sums, median, and mode). Aggregation of raster data always involves a cell size increase and a decrease in resolution. This is accomplished by multiplying the cell size of the input raster by a cell factor, which must be an integer greater than 1. For instance, a cell factor of 2 implies that the cell size of the output raster would be 2 times greater than cell size of input raster (e.g., an input resolution of 5km multiplied by 2 equals an output resolution of 10km). The cell factor also determines how many input cells are used to derive a value for each output cell. For example, a cell factor of 2 requires 2 × 2 or 4(2^2) input cells. The cell factor also determines how many input cells are used to derive a value for each output cell the following equation:

$$\textit{Output Cell Size} = \textit{Input Cell Size} \times \textit{Cell Factor}$$

In the proposed upscaling methodology, the value of each output cell is calculated as the mean or median of the input cells that fall within the output cell. In our study the scale factors will be 2, 5 and 10.

4. Methodology application of MEUSIS in Slovakia

The present chapter uses the results of a case study implemented in Slovakia in 2006 and the resulting Slovakia Soil Database. Due to financial resources, it is impossible to make such an

assessment on a larger scale and one of the EU-27 member states has been selected in order to perform the testing phase. In 2005-2006 period, the SSCRI, using its expertise to identify the appropriate local data sources, compiled the Slovakian Soil Database on three scales following MEUSIS requirements and, eventually, provided structured metadata as a complement part of the data. The data are considered relatively new in the soil science domain if you think that the European Soil Database contains national data which have been collected in the '70s and imported in digital format in the '80s.

Due to their specificity in terms of soil geography (variability in soil organic carbon content) and their data availability, the selected pilot areas in Slovakia have contributed to the analysis of the feasibility of such an innovative approach. In MEUSIS, all geographical information (Attributes and Geometry components) are represented by the grid of regular spatial elements (pixels). The representation of various spatial resolution details follows the INSPIRE recommendations. In addition, three **spatial resolution levels** of geographical information have been defined for MEUSIS:

- 10 km² (10km x 10km) coarse resolution grid, corresponding to data collection in national level
- 5 km² (5km x 5km) medium resolution grid, corresponding to data collection in regional level
- 1 km² (1km x 1km), fine resolution grid corresponding to data collection in local level

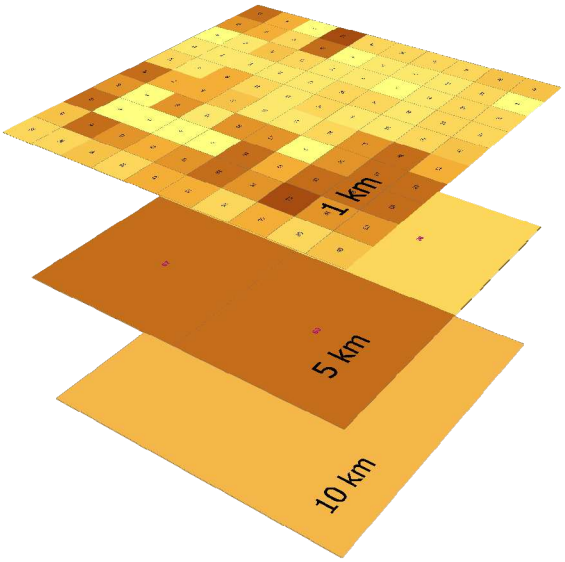


Fig. 2. Demonstration of upscaling

4.1 Upscaling from 5km² grid towards the 10km² grid

According to the aggregation technique described above, 4 cells of 5km x 5km size are requested in order to upscale their value to one single cell of 10 km x 10 km. The aggregation of the 5km x 5km grid cells is performed using both the MEAN value of the 4

cells and the MEDIAN value of the 4 cells producing 2 output datasets of 129 cells sized at10 km² each. In the cases near the borders, less than 4 cells are aggregated in order "produce" a cell of a coarser resolution at 10km².

The aggregation of 4 data cells using the Median function has an interesting drawback since if there are 3 cells out of 4 (cases near the borders of the input data) with 0 value, then the Median value of the 4 data cells is taking 0 value while the Mean value is different than 0. In order not to take into account those "extreme" cases which may alter our analysis, we will exclude the 5 cells. That implies that the 2 upscaled dataset plus the original one enclose 124 cells.

The present analysis may be applied also in order to identify cases where the data provider has previously performed the "tricky" operation well-known as downscaling. The proposed methodology can serve also as a first data quality check in order to find out if the data providers have contributed with their original data or they have manipulated their data by downscaling their coarser resolution data to finer resolution ones.

In figure 3, the scatter diagram reports the original 10km² values on the Y axis and the Upscaled (MEAN, MEDIAN) data on the Y axis. It is obvious that there is a noticeable linear relationship between the 2 upscaled datasets and the original data as there is a major concentration of data values near a line.

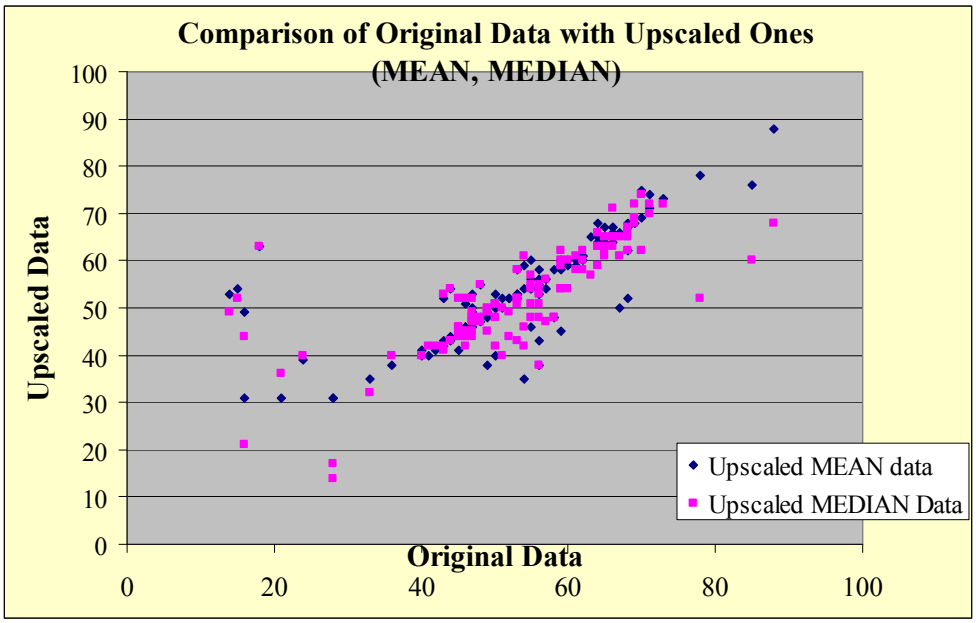


Fig. 3. Scatter Diagram of the Original data and Upscaled MEAN data

In the past, there were many theoretical references to an ideal MEUSIS as a nested system of hierarchical grids while in this analysis, we describe the results of the applied upscaling methodology in the Slovakian MEUSIS using both GIS operations and Statistical Analysis

(Descriptors, Scatter Diagram). Table 1 presents the core statistical indicators (Kavussanos, 2005) assessing the results of upscaling application.

Description of statistic	*Original Data 10km²*	*Upscaled data using MEAN*	*Upscaled data using MEDIAN*
Mean	52,96	53,76	52,29
Median	53	53	51,5
Mode	47	53	48
Standard Deviation	13,51	10,94	10,73
Sample Variance	182,61	119,58	115,18
Coefficient of Kurtosis	1,26	-0,03	1,30
Coefficient of Skewness	-0,65	0,25	-0,51
Range	74	57	60
Minimum	14	31	14
Maximum	88	88	74
P25 (First Quartile)	47	47	46
P75 (Third Quartile)	63	62	62
Count (Cells)	124	124	124
Confidence interval (95%)	2,40	1,94	1,91
Correlation Coefficient(r)		0,767	0,740

Table 1. Descriptive Statistics of the Upscaling Process from 5km² towards 10km²

The results of upscaling process which have used the MEAN value (named as Upscaled MEAN data) and the ones which have used the MEAN value (named as Upscaled MEDIAN data) will be compared against the Original data 10km² (supplied by the data provider) which is the criterion called to validate both processes. Find below the following remarks:

- The **Means** in both upscaled datasets are very close to the original data mean. Two are the possible explanations to this outcome:
  - Either the data sources for both the 10 km² Original and the 5 km² Original data are the same; this means that the original 5 km² numeric values, have previously been downscaled from the 10 km² Original ones. In practice, a newly introduced advantage of upscaling process is the detection of such data patterns. According to the data pattern, this is not the case in our datasets since the detailed data of 5 km² have a high variability inside the border of the 10km².
  - *Or* the use of the above mentioned upscaling method is producing satisfactory results.
- The **Median** values of both aggregated datasets are very close to the Median value of the original data. The **Mode** of upscaled MEDIAN data is very close to the mode of the original ones. Being almost the same, mean, median and mode of the upscaled MEAN data suggests symmetry in the distribution and once again confirm the theory that many naturally-occurring phenomena can be approximated by normal distributions (Dikmen, 2003).

Taking into account the three above mentioned measures of central tendency (Mean, Median, and Mode), we conclude that there are no extreme values that can affect the

distributions of the three datasets. There is a small-medium variability regarding the Organic Carbon Content in the scale of 5km² and as a consequence the upscaling process gives positive results either using the MEAN or the MEDIAN.

- **Range** and **Quartile indicators** show that there is quite medium variability in the original data which becomes smoother in the upscaled datasets.
- The original data have a relative higher **Standard Deviation** than the two upscaled datasets and it is evident that the two aggregated datasets show a "smooth" variability as they have reduced the dispersion of the data.
- **Data Distribution**: Regarding the prediction of intervals, it is it has been observed that the distribution of both upscaled data tends to be a normal distribution and as a consequence we may use the Standard Normal Distribution. With a **probability of 95%**, the range of possible values for the parameter Organic Carbon content 0-30cm will vary according to the equation;

$$P(-1.96\sigma \le X - \mu \le 1.96\sigma) = 0.95$$

All the above mentioned measures of dispersion show that upscaling process has a tendency for more smother data comparing with the original values.

- The frequency distributions in all three datasets are platykurtic (**Coefficient of Kurtosis)** and have a negative **Skewness** (except the original data with a symmetric distribution)
- **Correlation Coefficient or Pearson Correlation Coefficient** (r) is a measure of the strength of the linear relationship between two variables. It is not our objective to prove that there is a dependency between the 2 datasets; instead a possible high value of Coefficient indicates how good predictions we can make if we try to upscale the detailed data. The original 10km² data are used to **validate** how good forecasts can be given by the aggregated values. The value 0,767 determines a quite strong relationship between the upscaled MEAN data and the original ones (It is also obvious from the Scatter Diagram in Figure 3).

4.2 Upscaling from 1km² grid towards the 10km² grid

In order to update one cell of 10km x 10km, it is requested 100 cells of 1km x 1km. The data provider has collected data for 4.409 cells of 1km² which may be upscaled to 59 cells of 10km². In the cases near the borders, less than 100 cells are aggregated in order "produce" a cell of a coarser resolution at 10km. In Figure 4, the existing data covers only 14 1km² cells and the majority of the cells (11 out of 14) have 0 values. As a result the Mean is estimated with a value around 9 but the median will have a 0 value. In order not to take into account those "extreme" cases which may alter our analysis, we will exclude the 4 cells which have given results like the one shown above.

After implementing the upscaling process, the output datasets (Upscaled MEAN data, Upscaled MEDIAN data) have 55 common cells with the Original 10km² data. In the following paragraphs a more in depth statistical analysis will follow in order to assess the results of upscaling application.

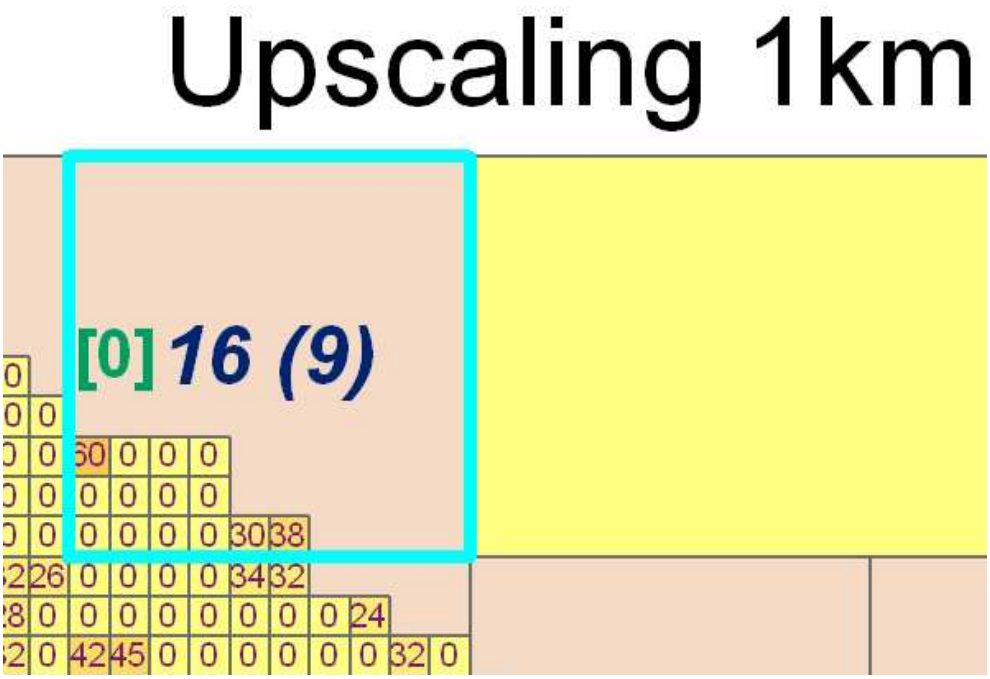


Fig. 4. The extreme case of MEDIAN upscale

Proceeding with the statistical analysis, some statistical descriptors are compared in the table 2 and the following remarks came out:

- Evaluating the **Mean** of the 3 datasets, we observe a slightly significant difference between the 2 Means of the upscaled data and the Mean of the original data. More than 10 tones per hectare difference may be explained as the upscaled data tend to have lower values than the original ones due to high dispersion of original data.
- Regarding the **Median** and the **Mode**, there is even a larger difference between the 2 upscaled datasets and the original data since the upscaling process has the trend to "produce" lower values.

Comparing the Upscaling results using the MEAN function with those using the MEDIAN function, we notice that the first ones tend to be better. The statistical indicators of the Upscaled MEAN data are closer to the Original data indicators. The upscaled MEDIAN data show a smoother dispersion and they show a big "concentration" around their mean.

- The **Range** of the Original data is higher than the one of the Upscaled MEAN data and much higher than the Upscaled MEDIAN data. The same comment is also referring to the P25 and P75 **Quartiles**.
- The **Standard Deviation** of the Upscaled MEAN data and the Original data are almost the same, while the standard deviation of the Upscaled MEDIAN data is much lower. The upscaled MEDIAN data show a very smooth variability while the other two datasets have almost the same variability.

- The **Correlation Coefficient** has a value of 0,49 between the Upscaled MEAN data and the Original data which express a medium-strong relationship (neither too strong, nor weak) between the 2 data distributions. Instead, this coefficient is smaller for the relationship between the Upscaled MEDIAN data and the Original ones which express a medium relationship between the 2 data distributions.

The results produced in the case of 1km² upscaling are considered satisfactory as the aggregation process that takes place aggregates 100 values to one. Scientists may argue that the upscale process may function well since averaging 100 values may "produce" a better result in an area of 10km² than picking up (survey) one random value in this large area (Original Data). At the end, comparing the upscaling results from 1km² with the ones from the 5km², we conclude that they are not as good as the latter ones. This remark can be explained since it is more probable to have good estimates when you upscale 4 cells than when you upscale 100 cells.

Description of statistic	*Original Data*	*Upscaled data using MEAN*	*Upscaled data using MEDIAN*
Mean	54,13	42,71	43,13
Median	56	40	44
Mode	50	29	30
Standard Deviation	15,43	16,00	10,56
Sample Variance	238,22	256,14	111,52
Coefficient of Kurtosis	1,29	1,11	0,20
Coefficient of Skewness	-0,77	0,98	0,06
Range	76	73	56
Minimum	12	16	16
Maximum	88	89	72
P25 (First Quartile)	47	33	35
P75 (Third Quartile)	65	52	50
Count (Cells)	55	55	55
Confidence Interval (95%)	4,17	4,33	2,85
Correlation Coefficient(r)		0,490	0,401

Table 2. Descriptive Statistics of the Upscaling Process from 1km² towards 10km²

4.3 Upscaling from 1km² grid towards the 5km² grid

In this case, the hierarchical grid system requests 25 cells of 1km² in order to update 1 cell of 5km². In the Slovakia Soil Database there are available 4.409 cells of 1km² and the upscaling process had as an output 207 cells of 5km². In this case, it was more evident the problem of the 0-value MEDIAN cells described above (with the Figure 4). In order not to alter the comparison results, the 20 cells with 0-value have been excluded and the outputs of 187 upscaled cells of 5km² will be compared in table 3.

Proceeding with the statistical analysis, some statistical descriptors are compared in the table 3 and the following remarks came out:

- The **Mean** values of the upscaled datasets are very close but still quite "distant" from the Mean value of the Original data. Around 8-9 tones per hectare difference may be explained as the upscaled data tend to have lower values than the original ones due to high dispersion of original data. Of course, the variability is less than the previous upscaling exercise since 25 cells is aggregated comparing with the 100 cells in the previous chapter.
- The **Standard Deviation** of the Upscaled MEAN data and the Original data are almost the same, while the Standard Deviation of the Upscaled MEDIAN data is much lower. The same "pattern" has been noticed in the previous upscaling exercise.
- The **Correlation Coefficient** has a value of 0,62 between the Upscaled MEAN data and the Original data which express a quite-strong relationship between the 2 data distributions. This indicator is used only to forecast how good can be possible predictions of the original data based on the upscaling processes.

Comparing the Upscaling results using the MEAN function with those using the MEDIAN function, we study that the first ones tend to follow the data pattern of the original data. Instead, the upscaled MEDIAN data show a smoother variability since they are more concentrated around their mean value. The statistical indicators, in the case of $1km^2$ upscaling towards $5km^2$, can be considered somehow in between the other 2 exercises with closer trend towards the results of the $1km^2$ to $10km^2$ upscaling. This remark can be explained since statistically it is more probable to have worst estimates when you upscale 25 cells than when you upscale 4 cells and better estimates than upscaling 100 cells.

Description of statistic	*Original Data*	*Upscaled data using MEAN*	*Upscaled data using MEDIAN*
Mean	54,98	46,21	45,75
Median	57	40	45
Mode	55	38	36
Standard Deviation	21,42	22,69	12,65
Sample Variance	458,82	514,97	160,12
Coefficient of Kurtosis	5,11	10,24	1,07
Coefficient of Skewness	-0,01	2,75	0,52
Range	161	154	84
Minimum	0	15	15
Maximum	161	169	99
P25 (First Quartile)	49	34	36
P75 (Third Quartile)	65	51	53
Count (Cells)	207	187	187
Confidence Interval (95%)	2,94	3,27	1,83
Correlation Coefficient(r)		0,62	0,54

Table 3. Descriptive Statistics of the Upscaling Process from $1km^2$ towards $5km^2$

5. Cross-comparison and conclusions on the 3 upscaling exercises

Major objective of this chapter is to analyse further the statistical indicators that have been described above, find out some more "interesting" relationships between various factors and compare the 3 upscaling exercises.

5.1 The "non-perfect squares" coverage effect

It has been observed in all three upscaling exercises that some squares have aggregated less input detailed data than required according to the Cell factor definition in the Technical Implementation. This observation is noticed in the borders of the data area. The concept of "**non-perfect squares**" is defined for those upscaled data cells where less than required data cells are aggregated.

In table 4, the Ration of Real to Expected squares can be defined as the percentage (%) of more cells that have been "produced" in the upscaling process due to the "non-Perfect Square" fact. In the first case there are 8,6% more cells than the expected ones, in the 1km^2 towards 5km^2 there are 17,4% more cells and in the 1km^2 towards 10km^2 upscaling there are 33,8% more cells. It is obvious that the Ratio of real to expected squares has a very strong positive relationship to the Cell Factor since it is increasing as the Cell Factor increases. Performing a regression analysis, the following outputs are found:

Ration = 1,02 + 0,031 * Cell Factor With coefficient of Determination: **R^2 = 0,9990**

Upscaling Exercise	*Cell Factor*	*Nr. of Input Cells*	*Expected squares (in case of perfect matching)*	*Real upscaled squares*	*Ratio of Real to expected*
5km towards 10km	2	475	118,75	129	1,086
1km towards 5km	5	4409	176,36	207	1,174
1km towards 10km	10	4409	44,09	59	1,338

Table 4. Analysis of "Non-Perfect Square"

The results are interesting allowing the modelers to identify how many more cells will have if they use an alternative Cell Factor. Even if this analysis may take different values in another country, the relationship between Cell Factor and additional cells will be always positive according to the "Non-Perfect Square" concept.

5.2 The role Correlation Coefficient (r) in predictions

Another interesting analysis can be considered the relationship between the Correlation Coefficient (r) in each of the 3 upscaling exercises with the Cell factor. In practice, this coefficient indicates how good can be the predictions given by the upscaling process validating them with the Original data.

In table 5, it is obvious that there is a negative relationship between the Correlation Coefficient (how good the predictions of upscaling can be) with the Cell Factor. As Cell Factor increases then the upscaling process will predict less precisely the real values.

Upscaling Exercise	*Cell Factor*	*Correlation Coefficient*
5km towards 10km	2	0,767
1km towards 5km	5	0,62
1km towards 10km	10	0,49

Table 5. Relation of Correlation Coefficient to Cell Factor

5.3 Lost of variation and dispersion variance

Commonly variation is lost when data are upscaled. This is modelled by the mean of the dispersion variance (Dungan et al, 2002) which quantifies the amount of lost variance between the 2 scales. Upscaling has a clear effect on spatial variability and this could be an advantage and disadvantage. In general for environmental data, if the interest focuses on observing extreme values in space, then upscaling is disadvantageous as the coarser scale variation tends to be smoother. But in case policy making involves recognition of general pattern then smoothing may be considered advantageous. We conclude that the latter is the case where soil organic carbon belongs to. The data variability or variance is smoothening since the upscaled values become smaller compared to the real finer scale data and this fact has been observed in all three upscaling exercises.

For comparison of the variability between the different sources, the coefficient of variation (Post el al, 2008) or the variances may be used. Alternatively, in the table 3, there is a comparison of the Variances, Ranges, Cell Factor, and Number of output cells between the 3 upscaling exercises. It is well known and it is proven in present case that variability is affected by the sample size and the extreme scores. The sample size is the number of output cells. It is supposed that variance should decrease as the number of output cells increases. This is not the case in the upscaled results because **the most important factor is the Range which determines the variance**. The high variability is due to the extreme values and as a consequence of the high ranges. This is proven in the orange part of the Table 3 and the trend of the variability in any of the 3 datasets (and upscaled exercises) is strongly affected by the trend of the Range in any direction of the table.

Upscaling Exercise	*Original data*	*Upscaled MEAN data*	*Upscaled MEDIAN data*	*Cell Factor*	*No of Output cells*
	Variance (Range)				
5 km² towards 10 km²	182,61 (74)	119,58 (57)	115,18 (60)	2	124
1 km² towards 10 km²	238,22 (76)	256,14 (73)	111,52 (56)	5	55
1 km² towards 5 km²	458,82 (161)	514,97 (154)	160,12 (84)	10	187

Table 6. Cross Comparison of Variance, Range, Cell Factor and No of Cells in Upscaling.

The dispersion of variance quantifies the amount of lost variance lost between scales. It is obvious from the table 3 that the median decreases the variance in upscaling.

5.4 Smoothing effect

Variation is lost when upscaling is performed. In case policy makers are interested in extremes values then upscaling has a disadvantage as either low or high values are smoothened. The smoothing effect is visible in figure 5 where the upscaled values have a smooth appearance. Instead the original 1km² values allow the policy maker to identify the extreme cases.

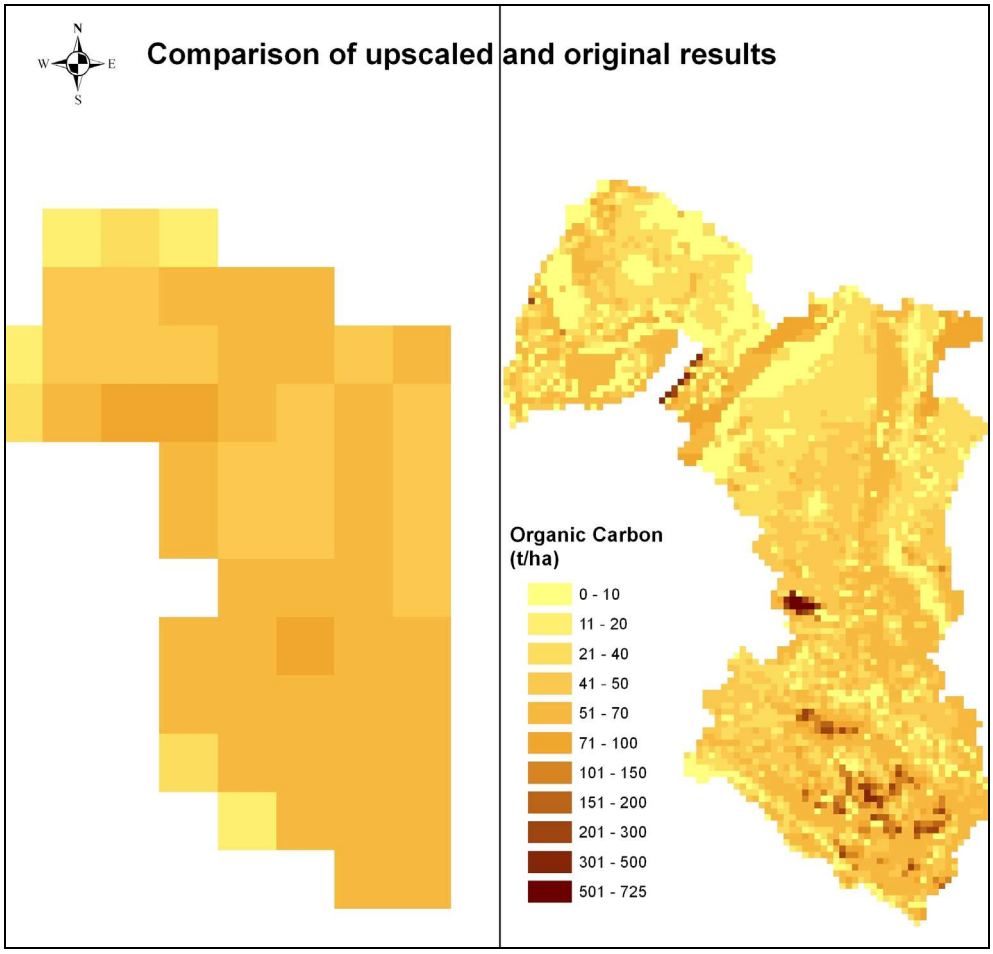


Fig. 5. The smooth effect in upscaling for the region Trnava in Slovakia

In case the policy maker is interested in the general pattern of the environmental indicator, then the upscaling proved to be advantageous. The advantage/disadvantage of upscaling depends also on the study area. In case the policy maker is interested in a small local region/province then the upscaled results may not be sufficient for his decision; instead in a larger scale (national), the identification of a pattern is much better succeeded with upscaled results than the raw data. Most of upscaled data are in the range between 51-70 t/ha C in the left part of the figure 5. In the majority of the cases, policy making is not based on the single observations but on general pattern. Instead a spatial study focusing in a specific area is disadvantageous using upscaled data. Comparison in time is better performed for the upscaled results since it allows the user to identify changes in block of cells.

Another reason for upscaling data is to ensure confidentiality during dissemination of data. This may be achieved by aggregated to various coarser scales than the size of data collection. European laws are quite strict in personal data treatment and land information data are quite sensitive and may affect the price of parcels. Suppose that you own an agricultural land parcel inside the1km^2 grid cell sample size and that information related to the sensitive environmental data (Organic carbon content, pH – Acidity, Heavy metal content, salinity…etc) about this cell are published. The parcel price is immediately affected by such publication and then the personal data protection authorities intervene and don't permit this kind of sensitive information dissemination. Instead, the process of data aggregation and the upscale of various environmental parameters in coarser scale make feasible the publication of low resolution land thematic maps without taking the risk of personal data violation. This implies that such a map must guarantee that individual entities (soil data) cannot be identified by users of the data. Aggregation is the traditional means for ensuring such confidentiality.

6. Spatial prediction and digital soil mapping

Digital Soil mapping (DSM) is the geostatistical procedure based on a number of predictive approaches involving environmental covariates, prior soil information in point and map form, (McBratney et al., 2003) and field and laboratory observational methods coupled with spatial and non-spatial soil inference systems (Carre et al., 2007). It allows for the prediction of soil properties or classes using soil information and environmental covariates of soil.

High-resolution and continuous maps are an essential prerequisite for precision agriculture and many environmental studies. Traditional, sample-based mapping is costly and time consuming, and the data collected are available only for discrete points in any landscape. Thus, sample-based soil mapping is not reasonably applicable for large areas like countries. Due to these limitations, Digital Soil Mapping (DSM) techniques can be used to map soil properties (Yigini et al., 2011).

As an example of the application of geostatistical techniques to produce continuous map of soil properties can be seen in the study conducted in Slovakia (Yigini et al., 2011). The authors studied to interpolation of point data to produce continuous map of soil organic carbon content in Slovakia. The regression kriging technique was applied and Corine Land

Cover 2006 data, SRTM 90m, European Soil Database (ESDB), climate, land management data were used as covariates. As a result, the soil organic carbon map was produced in raster format at a spatial resolution of 100 meters (Figure 6).

Digital Soil Mapping (DSM) can be defined as the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge and from related environmental variables (A.E. Hartemink et al., 2008). For soil mapping purposes, geostatistical techniques can be used to predict the value of the soil property at an unvisited or unsampled location by using auxiliary data (Figure 6). Most used interpolation methods are listed below;

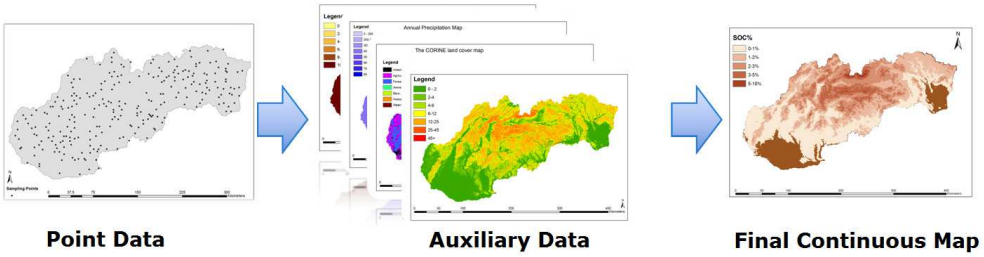


Fig. 6. Soil Properties can be mapped using geostatistical techniques

1. *Inverse distance weighting (IDW)*

Inverse Distance Weighted (IDW) is a technique of interpolation to estimate cell values by averaging the values of sample data points in the neighbourhood of each processing cell.

2. *Regularized spline with tension (RST)*

Regularized Spline with Tension (RST) is an accurate, flexible and efficient method for multivariate interpolation of scattered data (Hofierka et al., 2002)

3. *Ordinary kriging (OK)*

Ordinary Kriging is a geostatistical method used for regionalization of point data in space. Because it is similar to multiple linear regressions and interpolates values based on point estimates, it is the most general, widely used of the Kriging methods (Ahmed and Ibrahim, 2011)

4. *Ordinary co-kriging (OCK)*

Co-kriging allows samples of an auxiliary variable (also called the covariable), besides the target value of interest, to be used when predicting the target value at unsampled locations. The co-variable may be measured at the same points as the target (co-located samples), at other points, or both. The most common application of co-kriging is when the co-variable is cheaper to measure, and so has been more densely sampled, than the target variable (Rossiter, 2007)

5. Regression Kriging (RK)

Regression kriging is a spatial prediction technique which adds the regression value of exhaustive variables and the kriging value of residuals together (Sun et al., 2010).

7. Conclusions

On the basis of this study, the following conclusions can be drawn:

- The multi-scale nested grids approach can be proposed as a solution in many cases where the data owner does not allow the distribution/publication of detailed data but is willing to distribute degraded data (in coarser resolution). The aggregation methodology can be considered a valuable one which contributes to the degradation (without losing the real values) of very detailed data and may allow the scientific community to access valuable information without having any copyright problems.
- For a number of reasons upscaling can be useful in soil science domain: respect of privacy and data ownership, easy adaptation to model requirements, update of spatial databases in various scales and simplification of thematic maps.
- Upscaling methodology has proven to be good enough for identification of "data patterns". The upscaling process can easily identify if soil data have been downscaled before a possible aggregation for reporting reasons.
- Upscaling has a serious drawback in case the source dataset in the finer scale has high spatial variability. This has been shown in the upscaling process from 1km^2 towards the 10km^2. The descriptive statistics show the smooth effect that upscaling may cause in high variability cases. Upscaling involves recognition of general pattern in data distribution and this can be considered an advantage for environmental indicators. In any case the upscaled output doesn't represent the real world but it is a smooth approximation. The upscaling from local scale to upper scales involves trade-offs and compromises.
- Despite the limitations, the scale transfer method presented here was well-suited to the data and satisfied the overall objective of mapping soil indicators in coarser scale giving appropriate responses to policy makers. Moreover, a series of newly introduced concepts/indicators such as "Non-Perfect Square" Coverage, Correlation Coefficient for predictions and Lost of Variation can be introduced for further research and evaluation.
- Digital Soil Mapping (DSM) offers new opportunities for the prediction of environmental indicators in various scales.

8. References

Ahmed A., Ibrahim M., Production of Digital Climatic Maps Using Geostatistical Techniques (Ordinary Kriging) Case Study from Libya. International Journal of Water Resources and Arid Environments 1(4): 239-250, 2011 ISSN 2079-7079

Bechini L., Castoldi N., Stein A. Sensitivity to information upscaling of agro-ecological assessments: Application to soil organic carbon management (2011) Agricultural Systems, 104 (6), pp. 480-490.

Bierkens, M.F.P., Finke, P.A., De Willigen, P., 2000. Upscaling and Downscaling Methods for Environmental Research. Kluwer Academic Publishers, Dordrecht.

Cantelaube P., Jayet P.A., Carre F., Bamps C., Zakharov P. Geographical downscaling of outputs provided by an economic farm model calibrated at the regional level (2012) Land Use Policy, 29 (1), pp. 35-44.

Carre F., McBratney A.B., Mayr T., Montanarella L. Digital soil assessments: Beyond DSM (2007) Geoderma, 142 (1-2), pp. 69-79.

Dikmen, O., Akin, L., Alpaydin E., 2003. Estimating Distributions in Genetic Algorithms. Computer and Information sciences, Volume 2869/2003, pp. 521-528

Dungan, J. L., Perry, J. N., Dale, M. R. T., Legendre, P., Citron-Pousty, S., Fortin, M.-J., Jakomulska, A., Miriti M., and Rosenberg, M. S. 2002. A balanced view of scale in spatial statistical analysis. Ecography 25, pp. 626-640

EC, 2006. European Commission, 2006. Proposal for a Directive of the European Parliament and of the Council establishing a framework for the protection of soil and amending Directive 2004/35/EC. Brussels, 22-9-2006. COM (2006)232 final.

Ehleringer, JR and Field C.B 1993, Scaling physiological process: Leaf to globe, Academic Press, San Diego a.o, 388

Feddes, R.A., 1995. Space and Time Variability and Interdependencies in Hydrological Processes. Cambridge University Press, Cambridge.

Gaunt, J.L., Riley, J., Stein, A., Penning de Vries, F.W.T., 1997. Requirements for effective modelling strategies. Agric. Syst. 54, 153–168.

Hartemink, A.E., A.B. McBratney & L. Mendonca (eds) 2008 Digital soil mapping with limited data. Springer, Dordrecht. 445 pp. ISBN 978-1-4020-8591-8.

Harter, T. and J. W. Hopmans, 2004. Role of Vadose Zone Flow Processes in Regional Scale

Hydrology: Review, Opportunities and Challenges. In: Feddes,R.A., G.H. de Rooij and J.C. van Dam, Unsaturated Zone Modeling: Progress, Applications, and Challenges, (Kluwer, 2004), p. 179-208.

Hofierka J., Parajka J., Mitasova H.,Mitas L., Multivariate Interpolation of Precipitation Using Regularized Spline with Tension, Transactions in GIS (2002) Volume: 6, Issue: 2, Publisher: Wiley Online Library, Pages: 135-150 ISSN: 14679671

Hoosbeek, M.R., Bouma, J., 1998. Obtaining soil and land quality indicators using research chains and geostatistical methods. Nutr. Cycling Agroecosyst. 50, 35–50.

INSPIRE, 2007. INSPIRE EU Directive, Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), Official Journal of the European Union, L 108/1 50 (25 April 2007).

Jones, R.J.A., Hiederer, B., Rusco, F., Montanarella, L., 2005. Estimating organic carbon in the soils of Europe for policy support. European Journal of Soil Science 56, 655–671.

Kavussanos, M.G. and D. Giamouridis, Advanced Quantitative Methods for Managers Vol. 2 - Economic and Business Modelling. Hellenic Open University, Patras, 2005

King, A.W. 1991. Translating models across scales in the landscape, in Turner, M.G and Gardner, R.H(Eds), Quantitative methods in landscape ecology, Ecological studies, Springer, New York, Vol. 82, pp 479-517

Lal, R., 2004. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. Science 304 (2004), pp. 1623–1627.

McBratney, A.B., 1998. Some considerations on methods for spatially aggergatinbg and disaggregating soil information. Nutr. Cycling Agroecosyst. 50, 51–62.

McBratney, A., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

Minasny B., McBratney A.B. Spatial prediction of soil properties using EBLUP with the Matérn covariance function (2007) Geoderma, 140 (4), pp. 324-336.

Panagos P., Van Liedekerke M., Jones A., Montanarella L. European Soil Data Centre: Response to European policy support and public data requirements (2012) Land Use Policy, 29 (2), pp. 329-338, doi:10.1016/j.landusepol.2011.07.003, ISSN: 02648377

Post, J., Hattermann, F., Krysanova, V., Suckow, F., 2008. Parameter and input data uncertainty estimation for the assessment of long-term soil organic carbon dynamics. Environmental Modelling & Software. Volume 23, Issue 2, February 2008, Pages 125-138

Rossiter D. G., 2007Technical Note: Co-kriging with the gstat package of the R environment for statistical computing.

Rubin, Y., and D. Or. 1993. Stochastic modeling of unsaturated flow in heterogeneous media with water uptake by plant roots: The parallel columns model. Water Resour. Res. 29:619–631.

Saffih-Hdadi, K., Mary, B., 2008. Modeling consequences of straw residues export on soil organic carbon. Soil Biol. Biochem. 40, 594–607.

Stein, A., Riley, J., Halberg, N., 2001. Issues of scale for environmental indicators. Agric. Ecosyst. Environ. 87, 215–232.

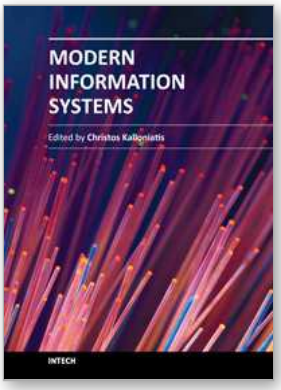
Sun W., Minasny B., McBratney A., Local regression kriging approach for analysing high density data. 19th World Congress of Soil Science, Soil Solutions for a Changing World 1 – 6 August 2010, Brisbane, Australia. Published on DVD

Tillotson, P.M., and D.R. Nielsen. 1984. Scale factors in soil science. Soil Sci. Soc. Am. J. 48:953–959

Viglizzo, E.F., Frank, F., Bernardos, J., Buschiazzo, D.E., Cabo, S., 2006. A rapid method for assessing the environmental performance of commercial farms in the Pampas of Argentina. Environ. Monit. Assess. 117, 109–134.

Van Bodegom, P.M., Verburg, P.H., Stein, A., Adiningsih, S., Denier van der Gon, H.A.C.: Effects of interpolation and data resolution on methane emission estimates from rice paddies. Environ. Ecol. Stat. 9(1), 5–26 (2002)

Yigini Y., Panagos P., Montanarella L., Spatial Prediction of Soil Organic Carbon Using Digital Soil Mapping Techniques in Slovakia - Volume 75, Number 3, June 2011, Mineralogical Magazine, Goldschmidt Abstracts 2011 - ISSN 0026-461X, Online ISSN: 1471-8022



Modern Information Systems

Edited by Dr. Christos Kalloniatis

ISBN 978-953-51-0647-0
Hard cover, 166 pages
Publisher InTech
Published online 13, June, 2012
Published in print edition June, 2012

The development of modern information systems is a demanding task. New technologies and tools are designed, implemented and presented in the market on a daily bases. User needs change dramatically fast and the IT industry copes to reach the level of efficiency and adaptability for its systems in order to be competitive and up-to-date. Thus, the realization of modern information systems with great characteristics and functionalities implemented for specific areas of interest is a fact of our modern and demanding digital society and this is the main scope of this book. Therefore, this book aims to present a number of innovative and recently developed information systems. It is titled "Modern Information Systems" and includes 8 chapters. This book may assist researchers on studying the innovative functions of modern systems in various areas like health, telematics, knowledge management, etc. It can also assist young students in capturing the new research tendencies of the information systems' development.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Panos Panagos, Yusuf Yigini and Luca Montanarella (2012). Use of Descriptive Statistical Indicators for Aggregating Environmental Data in Multi-Scale European Databases, Modern Information Systems, Dr. Christos Kalloniatis (Ed.), ISBN: 978-953-51-0647-0, InTech, Available from: http://www.intechopen.com/books/modern-information-systems/use-of-descriptive-statistical-indicators-for-aggregating-environmental-data-in-multi-scale-euro

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821