Geophysical Research Abstracts
Vol. 15, EGU2013-11171, 2013
EGU General Assembly 2013

Comparing different approaches - data mining, geostatistic, and deterministic pedology - to assess the frequency of WRB Reference Soil Groups in the Italian soil regions

Romina Lorenzetti, Roberto Barbetti, Giovanni L'Abate, Maria Fantappiè, and Edoardo A C Costantini

Consiglio per la ricerca e la sperimentazione in agricoltura. CRA-ABP Agrobiology and pedology research center, Florence, Italy

Estimating frequency of soil classes in map unit is always affected by some degree of uncertainty, especially at small scales, with a larger generalization.
The aim of this study was to compare different possible approaches - data mining, geostatistic, deterministic pedology - to assess the frequency of WRB Reference Soil Groups (RSG) in the major Italian soil regions.
In the soil map of Italy (Costantini et al., 2012), a list of the first five RSG was reported in each major 10 soil regions. The soil map was produced using the national soil geodatabase, which stored 22,015 analyzed and classified pedons, 1,413 soil typological unit (STU) and a set of auxiliary variables (lithology, land-use, DEM). Other variables were added, to better consider the influence of soil forming factors (slope, soil aridity index, carbon stock, soil inorganic carbon content, clay, sand, geography of soil regions and soil systems) and a grid at 1 km mesh was set up.
The traditional deterministic pedology assessed the STU frequency according to the expert judgment presence in every elementary landscape which formed the mapping unit.
Different data mining techniques were firstly compared in their ability to predict RSG through auxiliary variables (neural networks, random forests, boosted tree, supported vector machine (SVM)). We selected SVM according to the result of a testing set. A SVM model is a representation of the examples as points in space, mapped so that examples of separate categories are divided by a clear gap that is as wide as possible.
The geostatistic algorithm we used was an indicator collocated cokriging. The class values of the auxiliary variables, available at all the points of the grid, were transformed in indicator variables (values 0, 1). A principal component analysis allowed us to select the variables that were able to explain the largest variability, and to correlate each RSG with the first principal component, which explained the 51% of the total variability. The principal component was used as collocated variable. The results were as many probability maps as the estimated WRB classes. They were summed up in a unique map, with the most probable class at each pixel.
The first five more frequent RSG resulting from the three methods were compared.
The outcomes were validated with a subset of the 10% of the pedons, kept out before the elaborations. The error estimate was produced for each estimated RSG.
The first results, obtained in one of the most widespread soil region (plains and low hills of central and southern Italy) showed that the first two frequency classes were the same for all the three methods. The deterministic method differed from the others at the third position, while the statistical methods inverted the third and fourth position.
An advantage of the SVM was the possibility to use in the same elaboration numeric and categorical variable, without any previous transformation, which reduced the processing time.
A Bayesian validation indicated that the SVM method was as reliable as the indicator collocated cokriging, and better than the deterministic pedological approach.