Geophysical Research Abstracts
Vol. 19, EGU2017-18641, 2017
EGU General Assembly 2017

Selecting minimum dataset soil variables using PLSR as a regressive multivariate method

Anna Maria Stellacci (1,2), Elena Armenise (3), Mirko Castellini (1), Roberta Rossi (4), Carolina Vitti (1), Rita Leogrande (1), Daniela De Benedetto (1), Rossana M. Ferrara (1), and Gaetano A. Vivaldi (2)

(1) Council for Agriculture Research and Economics, Research unit for agriculture in dry environments (CREA-SCA), Bari, Italy (annamaria.stellacci@crea.gov.it), (2) University of Bari "Aldo Moro", DISSPA and DISAAT Departments, IT, (3) School of Energy, Environment and Agrifood, Cranfield University, UK , (4) Council for Agriculture, Research and Economics, Research unit for the extensive animal husbandry (CREA-ZOE), IT

Long-term field experiments and science-based tools that characterize soil status (namely the soil quality indices, SQIs) assume a strategic role in assessing the effect of agronomic techniques and thus in improving soil management especially in marginal environments.
Selecting key soil variables able to best represent soil status is a critical step for the calculation of SQIs. Current studies show the effectiveness of statistical methods for variable selection to extract relevant information deriving from multivariate datasets. Principal component analysis (PCA) has been mainly used, however supervised multivariate methods and regressive techniques are progressively being evaluated (Armenise et al., 2013; de Paul Obade et al., 2016; Pulido Moncada et al., 2014). The present study explores the effectiveness of partial least square regression (PLSR) in selecting critical soil variables, using a dataset comparing conventional tillage and sod-seeding on durum wheat. The results were compared to those obtained using PCA and stepwise discriminant analysis (SDA).
The soil data derived from a long-term field experiment in Southern Italy. On samples collected in April 2015, the following set of variables was quantified: (i) chemical: total organic carbon and nitrogen (TOC and TN), alkali-extractable C (TEC and humic substances – HA-FA), water extractable N and organic C (WEN and WEOC), Olsen extractable P, exchangeable cations, pH and EC; (ii) physical: texture, dry bulk density (BD), macroporosity (Pmac), air capacity (AC), and relative field capacity (RFC); (iii) biological: carbon of the microbial biomass quantified with the fumigation-extraction method.
PCA and SDA were previously applied to the multivariate dataset (Stellacci et al., 2016). PLSR was carried out on mean centered and variance scaled data of predictors (soil variables) and response (wheat yield) variables using the PLS procedure of SAS/STAT. In addition, variable importance for projection (VIP) statistics was used to quantitatively assess the predictors most relevant for response variable estimation and then for variable selection (Andersen and Bro, 2010).
PCA and SDA returned TOC and RFC as influential variables both on the set of chemical and physical data analyzed separately as well as on the whole dataset (Stellacci et al., 2016). Highly weighted variables in PCA were also TEC, followed by K, and AC, followed by Pmac and BD, in the first PC (41.2% of total variance); Olsen P and HA-FA in the second PC (12.6%), Ca in the third (10.6%) component. Variables enabling maximum discrimination among treatments for SDA were WEOC, on the whole dataset, humic substances, followed by Olsen P, EC and clay, in the separate data analyses.
The highest PLS-VIP statistics were recorded for Olsen P and Pmac, followed by TOC, TEC, pH and Mg for chemical variables and clay, RFC and AC for the physical variables.
Results show that different methods may provide different ranking of the selected variables and the presence of a response variable, in regressive techniques, may affect variable selection. Further investigation with different response variables and with multi-year datasets would allow to better define advantages and limits of single or combined approaches.

Acknowledgment
The work was supported by the projects "BIOTILLAGE, approcci innovative per il miglioramento delle performances ambientali e produttive dei sistemi cerealicoli no-tillage", financed by PSR-Basilicata 2007–2013, and "DESERT, Low-cost water desalination and sensor technology compact module" financed by ERANET-WATERWORKS 2014.

References

Andersen C.M. and Bro R., 2010. Variable selection in regression – a tutorial. Journal of Chemometrics, 24 728-737.

Armenise et al., 2013. Developing a soil quality index to compare soil fitness for agricultural use under different managements in the mediterranean environment. Soil and Tillage Research, 130:91-98.

de Paul Obade et al., 2016. A standardized soil quality index for diverse field conditions. Sci. Total Env. 541:424-434.

Pulido Moncada et al., 2014. Data-driven analysis of soil quality indicators using limited data. Geoderma, 235:271-278.

Stellacci et al., 2016. Comparison of different multivariate methods to select key soil variables for soil quality indices computation. XLV Congress of the Italian Society of Agronomy (SIA), Sassari, 20-22 September 2016.