Computers and Electronics in Agriculture 154 (2018) 120–133

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Automatic prediction of village-wise soil fertility for several nutrients in India using a wide range of regression methods

M.S. Sirsat, E. Cernadas, M. Fernández-Delgado*, S. Barro

Centro Singular de Investigación en Tecnoloxías da Información da USC (CiTIUS), Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, Campus Vida 15782, Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
Indian agriculture
Soil fertility index
Machine learning
Regression
Extremely randomized regression trees

ABSTRACT

In low quality soils, as in the Indian state of Maharashtra, a sustainable land management practice is very important to enhance the soil quality and to maintain proper values for several nutrients that are relevant for an optimal crop yield. The evaluation of a soil fertility index for these nutrients and for each geographical place allows to create maps of village-wise fertility indices which are very useful for fertility management. An automatic prediction of such fertility indices would be very important to reduce the amount of chemical measurements of nutrients to be performed in different cultivation lands. The current study develops the prediction of fertility indices for soil organic carbon and four important soil nutrients (phosphorus pentoxide, iron, manganese and zinc) using almost all the available regression methods, specifically a collection of 76 regressors which belong to 20 families, including neural networks, deep learning, support vector regression, random forests, bagging and boosting, lasso and ridge regression, Bayesian models and more. The best results are achieved by the extremely randomized regression trees (extraTrees), with which achieve an acceptable prediction accuracy (average squared correlations between 0.57 and 0.70), being also relatively fast. Other regressors with high performance are random forests and regularized random forest, generalized boosting regression model and epsilon-support vector regression.

1. Introduction

Agriculture is one of the most important economic fields in India, but urbanization and industrialization reduces the cultivable land. There is a need of increase the agricultural production in a sustainable way and without harming the environment. This requires to plan the soil fertility by supplying essential nutrients to the crop in sufficient amount and at right time for its best growth. Imbalances in soil quality reduce the crop health and lead to lower crop yield (Research Council, 1989), being a highly significant factor for achieving high crop production. Declining status of soil fertility and mismanagement of soil nutrients may be factors for food crises for the world's population (Gruhn et al., 2000). Generally, Indian soil fertility data are summarized for block and district level. In the view of changes in fertility levels, these data are useful for: (1) decision making about application of suitable amounts of fertilizers; and (2) policies of fertilizer distribution and consumption. The creation of maps for village-wise fertility indices and for several relevant nutrients would be very useful to compare levels of soil fertility among villages, and to make fertilizer recommendations specific for each village. Although villages themselves do not represent the physical and chemical attributes of the soil, these attributes are directly measured for each soil in order to estimate its fertility, so the geographical information about villages, and the spatial assessment of soil fertility, will be implicitly present through the soil attributes for the each specific land field. For the development of such fertility maps, much effort in terms of chemical analysis and time of specialized staff might be avoided if the direct measurement of the soil fertility through nutrient levels, for each village, might be replaced by an automatic, accurate enough, prediction. Most of the literature about prediction of soil parameters is based on the concept of pedotransfer function (PTF), which allows to describe mathematical relations among soil properties, using measurements to predict or to estimate certain soil parameters which are missing or whose measurement is time-consuming or expensive (Bouma, 1989; Pachepsky et al., 2015). The PTF can be formulated using data mining, exploration and machine learning regression methods. After reviewing the literature about PTFs, the objective of the current paper is to use regression techniques as PTFs that automatically predict the village-wise soil fertility indices for several relevant soil nutrients including organic carbon (*OC*), phosphorus pentoxide (P_2O_5), iron (*Fe*), manganese (*Mn*), and zinc (*Zn*), using data

* Corresponding author.
E-mail address: manuel.fernandez.delgado@usc.es (M. Fernández-Delgado).

https://doi.org/10.1016/j.compag.2018.08.003
Received 5 September 2017; Received in revised form 3 July 2018; Accepted 2 August 2018
0168-1699/

from the Marathwada region in the center of the Maharashtra state of India. This automatic prediction would be very useful for the creation of village-wise fertility maps, although the creation of such maps falls outside the scope of the current research due to the limited availability of detailed mapping data. On the other hand, spatial and temporal aspects of the problem which are relevant for the prediction of soil fertility, such as the distance between cultivation lands where data are acquired, agriculture over-exploitation of each particular land and year of data acquisition, are deferred for future studies. The paper is organized as follows: Section 2 analyzes previous works which use machine learning methods to predict soil parameters; Section 3 describes the calculation of village-wise soil fertility indices which are predicted in the current study; Section 4 describes the datasets and regression methods used for the prediction of fertility indices, and Section 5 discusses the results of the experimental work. Finally, Section 7 compiles the conclusions of this work.

2. Related work

Several studies applied machine learning techniques to solve soil problems in agriculture. Mucherino et al. (2009) provides a review of the methods used, among other objectives, to predict the soil fertility, defined as its ability to supply the required nutrient levels and water for high quality crop yield. The soil organic carbon is predicted by Ritz et al. (2015) using an unbiased linear predictor. Schillaci et al. (2017) predicts the organic carbon on Sicilian soils applying boosted regression trees, which are also used by Wang et al. (2018), alongside with random forest, combined with genetic algorithms for feature selection, for the same task on soils of eastern Australia. Other methods used for mapping and predicting soil carbon are reviewed by Minasny et al. (2013). Soil fertility was predicted using neural networks with Levenberg-Marquadt based back-propagation (Sheela and Sivaranjani, 2015) and partial least squares regression (Obade and Lal, 2016), whose inputs included soil bulk density, electrical conductivity (*EC*), available water capacity, soil *OC*, pewamo silty clay loam, glynwood silt loam, kibbie fine sandy loam, crosby silt loam and crosby celina silt loams soil. Terhoeven-Urselmans et al. (2010) predicted acidity (*pH*), alongside with *OC* and cation exchange capacity from mid-infrared spectra for several soils using partial least-squares regression and the prediction root mean square error as quality measure. Jia et al. (2010) applied a Bayesian network for soil fertility grading using the soil *pH* and nutrients as copper (*Cu*), *Fe*, potassium (*K*), *Mn*, nitrogen (*N*), phosphorus (*P*), *OC* and *Zn*. Lamorski et al. (2008) found that support vector machine (SVM) outperforms neural networks to provide a PTF which predicts the soil total nitrogen using bulk density and soil contents of water, silt, sand and clay. Jain et al. (2004) focused on PTFs for the prediction of water retention and saturated/unsaturated hydraulic conductivity, properties which are expensive to measure. Minasny et al. (1999) used multiple-linear regression, extended nonlinear regression and neural networks to estimate water-retention PTFs for soils in Australia. The geographical transferability of machine learning algorithms has been assessed by Veronesi et al. (2017) for the modelling of wind speed. Heung et al. (2016) reviews machine learning techniques for the classification of soil into 11 orders and 18 great groups from satellite images at 100 m. spatial resolution, comparing classification and regression trees (CART), bagging ensemble of CART, random forest, k-nearest neighbors, logistic model tree, multinomial logistic regression, nearest shrunken centroid, neural network and SVM. In a previous study (Sirsat et al., 2017), we also used a collection composed by 20 classifiers, including random forests, neural networks, adaboost, SVM and bagging, among others, to classify several nutrient levels and village-wise soil fertility indices. The class labels were quantified values (low, medium and high) of their numeric values. We also classified the soil type and *pH*, and the recommended crop for the next cycle. In the current paper, we use an even larger and more diverse collection of regression methods in order to create PTFs which directly predict, without discretization, the numeric values of fertility indices for several important soil nutrients which will be described below.

3. Prediction of village-wise soil fertility indices

The soil of Marathwada is intensively cultivated with good agricultural practices such as organic farming, irrigation plannings, use of precision agriculture tools and soil health card system in order to increase crop yield. A major factor for soil productivity is fertility, which primarily deals with ability of soil to supply nutrients to plants. Fertility of agricultural soil is depleting due to intensive cultivation practices and inadequate or excessive use of chemical fertilizers. To attenuate these soil problems, there is a need of knowledge about soil physical and chemical status. The village-wise fertility indices for soil *OC*, P_2O_5 and micro (*Fe*, *Mn*, *Zn*) nutrients are not only helpful to choose the right fertilizer and dose, but also to know about inherent excess and deficiency in them, and to balance nutrients up to critical levels. The *OC* is very relevant for the biological activity of the soil and for crop productivity (Reeves, 1997), while P_2O_5 is necessary for cell signaling, phosphorylation and bioenergetics in plants. On the other hand, *Fe* and *Mn* are used by chlorophyll during photosynthesis to absorb energy from light. Finally, *Zn* contributes to the production of plant growth hormones and proteins, being responsible for plant root development as well as carbohydrate and chlorophyll formation (Arunachalam et al., 2013). The *Zn* affects the crop yield and soil quality, being the most deficient micro-nutrient in Indian soils by nearly 50% of the required amount (Intl. Zinc Assoc, 2018). The agriculture planning of the Indian Government requires to determine the village-wise soil fertility indices (denoted as N_I) for the previous nutrients and to quantify their levels as low, medium and high. Inspired by the previous ideas, the present work applies a collection of regression techniques to automatically predict village-wise soil fertility indices for the previous nutrients using several chemical parameters of the soil. Rammoorthy and Bajaj (1969) defined the procedure to calculate the village-wise soil fertility index for a nutrient. First, each cultivation lands in the village is evaluated, according to its fertility for the corresponding nutrient, as low, medium or high using the limits listed in Table 1. Second, the numbers N_L, N_M and N_H of cultivation lands with low, medium and high fertility levels, respectively, for the nutrient, and the total number of lands $N_T = N_L + N_M + N_H$, are determined for each village. Finally, the village-wise fertility index N_I is calculated using the following formula:

$$N_I = \frac{N_L + 2N_M + 3N_H}{N_T} \tag{1}$$

The value of N_I is a weighted average of the numbers of cultivation lands with low, medium and high fertility indices, so its value is restricted to the range $1 \leqslant N_I \leqslant 3$. Values of N_I near to 1 mean that low fertility fields are predominant for that nutrient and village; N_I values about 2 are associated to medium fertility index; and N_I values about 3 correspond to high fertility indices. The index value is the same for all the patterns in the village, whose cultivation lands share the same fertility index for every soil nutrient.

Table 1
Intervals defined by the Department of Agriculture & Cooperation (2011) of the Indian Government for the organic carbon (measured in %), phosphorus pentoxide (in kg/ha) and micro nutrients (in parts per million, PPM), respectively (Muhr et al., 1965; Katyal and Rattan, 2003).

	OC (%)	P_2O_5 (kg/ha)	*Fe* (PPM)	*Mn* (PPM)	*Zn* (PPM)
Low <	0.5	10	1	2.5	108
Medium	0.5–0.75	10–24.6	1–2	2.5–4.5	108–280
High >	0.75	24.6	2	4.5	280