Classification of agricultural soil parameters in India

Article *in* Computers and Electronics in Agriculture · April 2017

DOI: 10.1016/j.compag.2017.01.019

CITATIONS
19

READS
6,576

4 authors, including:

Manisha Sirsat
NOVA.ID.FCT
10 PUBLICATIONS **83** CITATIONS
SEE PROFILE

E. Cernadas
University of Santiago de Compostela
51 PUBLICATIONS **2,122** CITATIONS
SEE PROFILE

Razaullah Khan
Maulana Azad National Urdu University
12 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Classification of agricultural soil parameters in India

M.S. Sirsat[a], E. Cernadas[a], M. Fernández-Delgado[a,*], R. Khan[b]

[a] *Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS) Universidade de Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez Campus Vida 15782, Santiago de Compostela, Spain*
[b] *Department of Commerce & Management Science, Maulana Azad College, 431005, Aurangabad (M.S.), India*

Abstract

One of the backbones of the Indian economy is agriculture, which is conditioned by the poor soil fertility. In this study we use chemical soil measurements to classify many relevant soil parameters: village-wise fertility indices of organic carbon (OC), phosphorus pentoxide (P_2O_5), manganese (Mn) and iron (Fe); soil pH and type; soil nutrients nitrous oxide (N_2O), P_2O_5 and potassium oxide (K_2O), in order to recommend suitable amounts of fertilizers; and preferable crop. To classify these soil parameters allows to save time of specialized technicians developing expensive chemical analysis. These ten classification problems are solved using a collection of twenty very diverse classifiers, selected by their high performances, of families bagging, boosting, decision trees, nearest neighbors, neural networks, random forests (RF), rule based and support vector machines (SVM). The RF achieves the best performance for six of ten problems, overcoming 90% of the maximum performance in all the cases, followed by adaboost, SVM and Gaussian extreme learning machine. Although for some problems (pH, N_2O, P_2O_5 and K_2O) the performance is moderate, some classifiers (e.g. for fertility indices of P_2O_5, Mn and Fe) trained in one region revealed valid for other Indian regions.

Keywords: Soil type, machine learning, random forest, soil fertility, fertilizer recommendation.

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

1. Introduction

According to data of year 2011, India devotes 60.5% of its land[1] to agriculture (CIA, 2016), distributed among arable land (52.8%), land for permanent crops (4.2%) and pastures (3.5%). Share of agriculture and related activities was 11.3% of the Gross State Domestic product (GSDP) in 2013-14. Data from the Directorate of Economics and Statistics (2015) show that in year 2013-2014 the cultivation areas of major crops were 15 and 57 millions of hector in Kharif and Rabi seasons, respectively. However, agriculture in India is conditioned by the poor fertility of the soil, which depends on the levels of its nutrients. The physical, chemical and biological properties of the soil are useful to evaluate its fertility, to design a cultivation plan and to predict the crop productivity. The information technologies, and specifically machine learning (ML), offer new possibilities in the field of agriculture and may help in data evaluation for decision making. The geographical study area of the current paper is the region of Marathwada, in the state of Maharashtra, one of the most prominent agricultural states in India, located at 19^o 52' 59.88" North and 75^o 19' 59.88" East. Its soil is made of basalt rock with scarlet, blackish and yellowish colors. The soil classification is useful to maintain and enhance its productivity, to avoid soil degradation problems and to overcome environmental damage. The major challenge is to increase crop yield for solving global food security problem. However, soil quality and crop yield are negatively affected by changing trends of temperature and rainfall, insufficient water and light, agriculture practices and absence of nutrients. It is important to develop an effective nutrient management by means of an adequate soil analysis and a proper application of fertilizers. Hence the relevance of a research effort to classify soil parameters such as the fertility indices for several nutrients (OC, P_2O_5, Mn and Fe, among others), soil pH, soil type, preferred crop and levels of several nutrients (N_2O, P_2O_5 and K_2O)

[1] https://www.cia.gov/library/ publications/the-world-factbook/fields/2097.html

which are relevant for fertilizer recommendation. The interest of predicting the levels of these magnitudes with machine learning techniques is to avoid the need to chemically measure these magnitudes, thus reducing the cost of the analysis and saving time of specialized technicians. The current study tries to enhance the accuracy of soil problem interpretation for Indian agriculture, although similar studies would benefit other nations around the globe.

2. Related work

Several studies (Mucherino et al., 2009) have applied ML techniques to solve soil problems in agriculture, namely to predict soil fertility, defined as the soil ability to supply the required nutrient levels and water for high quality crop yield. The soil fertility was predicted using artificial neural networks (ANNs) with Levenberg-Marquadt based back-propagation (Sheela and Sivaranjani, 2015), and also using partial least squares regression (Obade and Lal, 2016) using as input data the 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. The crop yield has been predicted using clustering techniques (Narkhede and Adhiya, 2014). One-R, J48, K-nearest neighbors (KNN) and A priori classifiers have been used to predict wheat yield (Romero et al., 2013) using as inputs phenotypic plant traits (thousand grain weight, plant height, peduncle length, harvest index, spikelets number, grain number, grain weight and spike fertility). The wheat yield has also been quantified (Pantazi et al., 2016) as low, medium and high with supervised Kohonen and counter-propagation neural networks, and with XY-fusion models using multi-layer soil data: pH, moisture content, total nitrogen, total carbon, magnesium, calcium, cation exchange capacity, available phosphorus and satellite imagery crop growth characteristics. Decisions about insecticide application (either spray or non-spray) for leafroller pest monitoring on kiwifruit are recommended (Hill et al., 2014) using decision tree (DT), naive Bayes classifier, RF, adaboost, SVM and logistic regression (LR). The gener-

alized regression neural network was used to forecast plant diseases (Chtiouia et al., 1999) for leaf wetness prediction. The RF provides the best accuracy for mapping the soil class spatial distribution in three semi-arid study areas with different sets of environmental covariates (Brungard et al., 2015), compared to clustering algorithms, discriminant analysis, multinomial logistic regression (MLR), ANN, DT and SVM. The soil has also been classified in 11 orders and 18 great groups from satellite images at 100m spatial resolution, using classification and regression trees (CART), bagging with CART base classifiers, RF, KNN, nearest shrunken centroid, ANN, MLR, logistic model trees and SVM (Heung et al., 2016).

Our work deals with several problems analysed in some of the previous papers, such as prediction of soil type (although with a different set of classes) and fertility indices (which in our paper are specific for OC, P_2O_5, Fe and Mn). However, our study analyses more soil problems than the previous works, including nutrients N_2O, P_2O_5 and K_2O, which allows to develop a fertilizer recommendation, the soil pH and prediction of suitable crops. We also evaluate the validity of the classifiers trained in one region to model data from different regions. However, available data for the current paper do not allow to predict crop yield, insecticide application nor plant diseases, as in some previous papers. On the other hand, our data are exclusively chemical measurements (see section 3), excluding data such as satellite images and phenotypic plant traits. Finally, we use a wider and more diverse collection of classifiers than previous papers, which are specifically selected due to the good performance that they exhibited in the experimental comparison (Fernández-Delgado et al., 2014).

3. Materials and methods

In the current research we use data collected from the region of Marathwada by the State Government of Maharashtra (India) during years 2011 to 2015.

Details about calibration of each input magnitude are publically available[2]. The inputs that we use are the following soil parameters: N_2O, measured by the soil testing laboratories using alkaline permanganate (Jones, 1912; Subbaiah and Asija, 1956); OC, using carbon spectrophotometric (Bowman et al., 1991); pH, using pH meter method (Richards et al., 1954); EC, using EC meter method (Richards et al., 1954); K_2O, using flame photometric (Ford, 1954; Jackson, 1958); and P_2O_5, using the Olsen's method (Olsen, 1954). Micro nutrients as Fe, copper (Cu), zinc (Zn), Mn and boron (B), which are useful to evaluate imbalances in soil nutrients, are measured using atomic absorption spectroscopy (Leggeit and Argyle, 1983). The pH is expressed as the decimal logarithm of the Hydrogen concentration. The values of N_2O, K_2O and P_2O_5 are expressed in kilograms per hector (Kg/ha), while Fe, Cu, Zn, Mn, B and SO_4 are expressed as parts per million (PPM). The EC is expressed in mili-Siemens per centimeter (mS/cm), while OC and $CaCO_3$ are expressed as mass percentages (denoted as %).

Table 1: Intervals defined by the Department of Agriculture & Cooperation (2011) for the major and micro nutrients respectively (Muhr et al., 1965; Katyal and Rattan, 2003), and rate of nutrient index (Rammoorthy and Bajaj, 1969).

	Major nutrients		Micro nutrients		Index
	OC	P_2O_5	Mn	Fe	
	(%)	(Kg/ha)	(PPM)		
Low <	0.5	10	1	2.5	1.67
Medium	0.5-0.75	10-24.6	1-2	2.5-4.5	1.67-2.33
High >	0.75	24.6	2	4.5	2.33

[2]http://mpkv.ac.in/Handler2.ashx?sel=Data&&tbl=DataMst&&whr=Type=%27PDF%27%20and%20Id=2004 (in Marathi language).

3.1. Classification problems

The following subsections describe the ten soil parameters which are classified in the current work.

3.1.1. Soil OC, P_2O_5, Mn and Fe village-wise fertility indices

There are proofs of the interconnection among organic matter, ecosystem sustainability and soil fertility (Feller et al., 2012), which is important for crop yield. This fertility mainly depends on OC, N_2O, P_2O_5 and K_2O, considered soil major nutrients because they appear in large quantities, and also on micro nutrients Fe, Mn, Zn and Cu, which appear in smaller quantities. However, our work is restricted to fertility levels of OC, P_2O_5, Fe and Mn, due to data availability. The OC is very important for the soil health, biological activity and crop productivity (Reeves, 1997), and adequate fertilizers help to keep its level (Turmel et al., 2015). The P_2O_5 is used by plants for cell signaling, phosphorylation and bioenergetics, while Fe and Mn help chlorophyll to absorb light energy for photosynthesis. The agriculture planning of the Indian Government requires to determine the village-wise fertility indices N_I for the previous nutrients, using the thresholds listed in Table 1 to quantify their levels as low, medium and high. For each village and nutrient, N_l, N_m and N_h are the number of patterns (i.e., cultivation lands) with low, medium and high levels. The village-wise fertility index N_I for a nutrient is calculated as $N_I = (N_l + 2N_m + 3N_h)/N_t$, being N_t is the total number of patterns analysed for a village. The value of N_I (which is the same for all the patterns in the village) is then quantified into low, medium and high levels, according to the threshold values listed in the rightmost column of Table 1. The classification of the village-wise fertility index uses the inputs listed in the first four lines of Table 2. The labels OC-F, P_2O_5-F, Mn-F and Fe-F mean village-wise fertility index of OC, P_2O_5, Mn and Fe, respectively, whose values are completely different to the inputs OC, P_2O_5, Mn and Fe. Our data (see Table 2) only include patterns with OC-F and Fe-F (resp. Mn-F) in levels low and medium (resp. medium and high).

Table 2: Inputs, number of total patterns (#Patterns) and patterns per class for each classification problem. The labels SA, N, SAL and MAL mean slightly acid, neutral, slightly alkaline and moderately alkaline, respectively. The compounds SO_4 and $CaCO_3$ are sulfate and calcium carbonate respectively. The symbol — means that patterns of that class are not available.

Problem	Inputs	#Patterns	Classes and #patterns per class			
			Low	Medium	High	
OC-F	EC, OC, N_2O,	372	203	169	—	
P_2O_5-F	P_2O_5, K_2O,SO_4,	372	104	240	28	
Mn-F	Cu,Fe, Mn,	367	—	183	184	
Fe-F	Zn, B	372	276	96	—	
N_2O	EC, OC, P_2O_5, K_2O, SO_4, Cu, Fe, Mn, Zn, B	372	124	124	124	
P_2O_5	EC, OC, N_2O, K_2O, SO_4, Cu, Fe, Mn, Zn, B	372	124	124	124	
K_2O	EC, OC, N_2O, P_2O_5, SO_4,Cu, Fe, Mn, Zn, B	372	124	124	124	
pH	P_2O_5, K_2O, EC, OC, $CaCO_3$, Cu, Mn, Zn, Fe	1137	SA	N	SAL	MAL
			22	432	544	139
Crop	P_2O_5, K_2O, pH, EC, OC	2878	Bajra(R)	Cotton(I)	Cotton(R)	Soybean(R)
			185	324	712	1657
Soil	P_2O_5, K_2O, pH, EC, OC	1692	Light	Medium	Heavy	
			482	1210	—	

3.1.2. Classification of soil nutrients N_2O, P_2O_5 and K_2O

The direct measurement of soil N_2O is difficult, but it is largely present in the OC form (97-99%), so that it can be determined indirectly from the OC, e.g. using linear regression (Rashidi and Seilsepour, 2009). High N_2O levels, as in tropical agro-ecosystems, have negative effects on water, air, ecosystem and human health, limiting the crop growth (Groenigen et al., 2015) and ecosys-

tem productivity, which also depends on temperature, precipitations and atmospheric CO_2. The deficiencies of the Indian soils with respect to N_2O lead to a strong application of suitable fertilizers. The P_2O_5 is also very important for the soil fertility, as we explained in the previous subsection. The K_2O is involved in crop physiological functions, being strongly deficient in the Indian soils (Naidu et al., 2011), and the corresponding fertilizer (muriate of potash) is a non-renewable resource which can not be synthesized from other chemicals, so it can not be managed in large amounts. The classification of N_2O, P_2O_5 and K_2O is very useful because N, P and K are the most responsible nutrients for fertilizer recommendation. However, as we mentioned in the previous subsection, using the national limits defined by the Indian Government (Table 1) the available data for N_2O and K_2O only include patterns of one class. Therefore, we defined the limits in order to have equally populated low, medium and high classes. These classifications use the 10 inputs listed in the lines 5-7 of Table 2, and their results are used to recommend the suitable amounts of fertilizers in the following way. Let $C(X)$ be the predicted class for nutrient X, defined as $C(X)$= 1, 2 or 3 for classes low, medium and high, respectively, being the nutrient X=N_2O, X=P_2O_5 or X=K_2O. Let also $R(Y)$ be the amount, in kg/ha, of fertilizer Y, being Y=uria, Y=super phosphate or Y=muriate of potash the recommended fertilizers to correct the level of nutrients N_2O, P_2O_5 or K_2O, respectively. The Mahatma Phule Agricultural University (2016) recommends an amount $R(X, Y)$ of fertilizer Y to correct the level of nutrient X calculated using $C(X)$ and a pre-defined reference amount $F(Y)$ of fertilizer Y, according to an expression proposed by the technicians of the State Government of Maharashtra (see footnote in section 3):

$$R(X, Y) = \frac{6 - C(X)}{4} F(Y) \tag{1}$$

Therefore, $R(1, Y) = 5F(Y)/4$, $R(2, Y) = F(Y)$ and $R(3, Y) = 3F(Y)/4$ for $C(X)$=1, 2 and 3, respectively, so the recommended amount is 125%, 100% or 75% of $F(Y)$ when the level of nutrient X is low, medium and high, respectively.

3.1.3. Classification of soil pH

The pH is the scale of soil acidity or alkalinity, which affects the crop yield and all the soil parameters, because soil acidity is one of its major degradation problems. Specifically, the region of Marathwada has slightly alkaline soil (high pH), which leads to nutrient deficiency, low OC and high $CaCO_3$ levels, limits the crop growth and reduces the crop yield. The pH classification uses the inputs listed in the line 8 of Table 2. Although usually nine pH levels are considered (Figure 1), for our purpuses it is enough to discriminate between the four middle classes: slightly acidic (labeled SA), neutral (N), slightly alkaline (SAL) and moderately alkaline (MAL). The classification of pH into levels is useful to decide suitable crops and pesticides, and to evaluate microbial activity, nutrient levels and soil corrosivity.

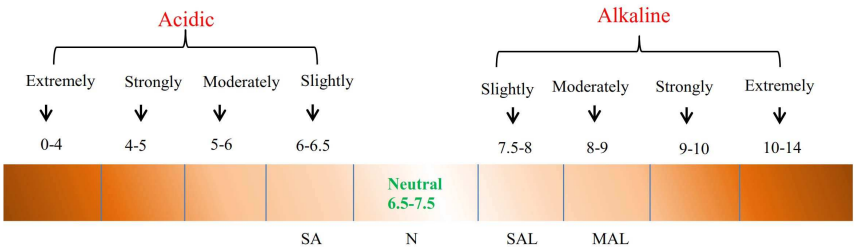

Figure 1: Degree of acidity and alkalinity of soil (Chesworth, 2008), with nine classes (up) and with the four classes considered in this paper (down).

3.1.4. Crop classification

The growth of a given crop needs a balanced supply of important nutrients, whose levels define the best crop for a given soil. The cropping cycle determines the way in which the soil parameters are enhanced: a good cycle is important for optimum yield and improvement of soil quality, erosion and moisture, organic matter and carbon storage (Jakubauskas et al., 2002). There are several studies about crop classification (Tatsumi et al., 2015), crop yield forecasting (Mkhabela

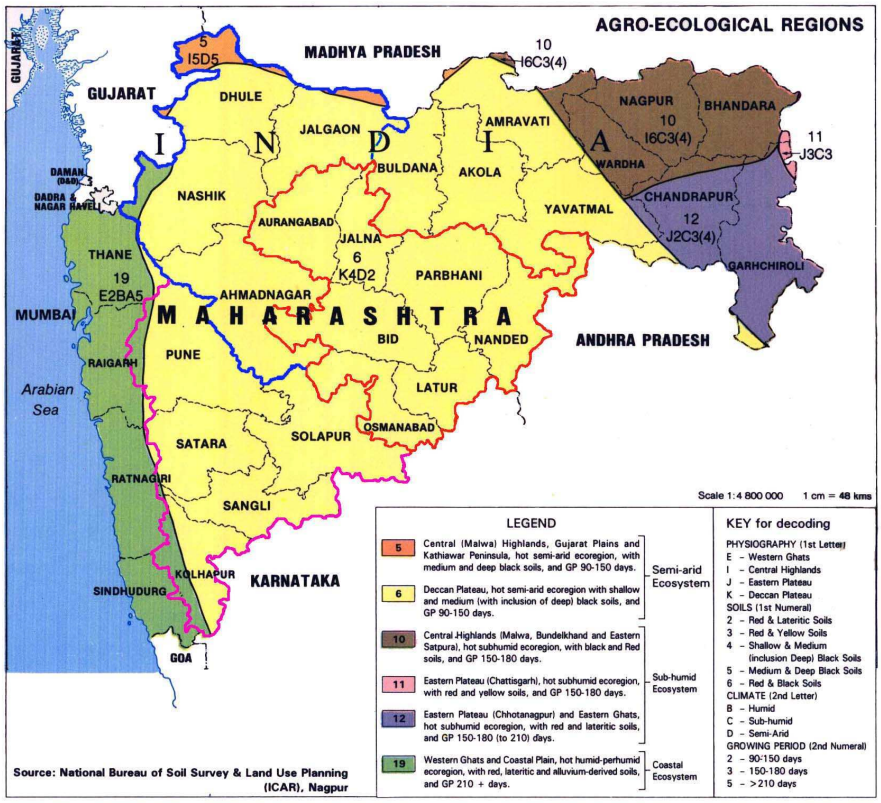

Figure 2: Soil map of Maharashtra (Sehgal, 1990). The geographical study areas are highlighted with outlines: red (Marathwada), blue (North Maharashtra) and pink (Paschim Maharashtra).

et al., 2011), mapping for crop rotation (Panigrahy and Sharma, 1997) and land cover classification on remotely sensed data using time-series analysis techniques (Rogan et al., 2008). The RF, linear discriminant analysis and SVM were used to classify crops (walnut, table grape, almond and European plum) on four feature sets (Pea and Brenning, 2015). The SVM was also used for plant discrimination using satellite land images (Guerrero et al., 2012). Our data includes four crops, which require neutral pH (in the range 6.5-7.5): bajra(R), cotton(I), cotton(R) and soybean(R), where R and I mean rainfed and irrigated, respectively. The Figure 3 (left panel) shows a geographical plot of the crop data, which are

distributed across a wide area including the Auangabad, Jalna, Bid, Osmanabad and Latur districts of the Marathwada region, inside of the state of Maharashtra. In the current paper, crop classification predicts which crop is suitable for the next stage of the cropping cycle using the inputs listed in the line 9 of Table 2.

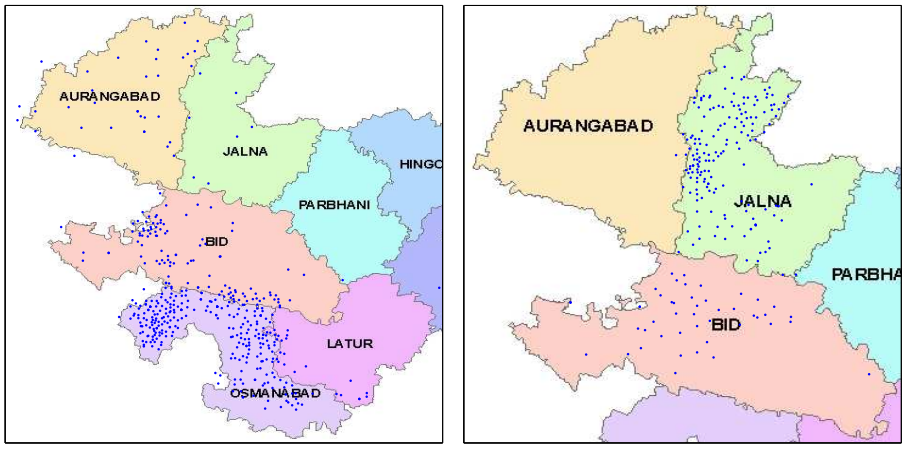

Figure 3: Maps with the geographical location of crop (left panel) and soil type (right panel) data over several districts in the Marathwada region (red outline in the map of Figure 2). Each point locates a different village, from where several patterns are recorded.

3.1.5. Soil classification by type

The soil classification according to its type allows to select the best soil for a particular crop. Soil has been classified (Taghizadeh-Mehrjardi et al., 2015) using LR, ANN, SVM, KNN, RF and DT in 5 types: 1) coarse loamy, mixed, mesic, lithic xerorthents; 2) fine, mixed, mesic, typic talcixerepts; 3) fine loamy, carbonatic, mesic, typic calcixerepts; 4) fine loamy, mixed, mesic, typic haploxerepts; and 5) fine, mixed, mesic, typic haploxerepts. The study used 217 patterns collected from Kurdistan Province, North-West Iran, achieving accuracy and Cohen kappa of 71% and 69% using DT and ANN, respectively. Figure 2 shows[3] the different types of soil in the state of Maharashtra (label

[3]http://eusoils.jrc.ec.europa.eu/esdb_archive/EuDASM/Asia/images/maps/download/in3010_3so.jpg

6 locates the soil type of Marathwada). In our study, three soil classes are considered according to its texture. Light soil has large proportion of sand, low parameter levels and ability to hold water. Sandy, peaty and chalky soils are subtypes of light soil. Medium (loam) soil contains silt, clay and humus (decayed matter) and it is suitable for several crops, being the predominant type in Marathwada. Heavy soil contains more moisture and sticky lump, due to the high proportion of silt (slightly larger particles of rock) or clay (small particles of rock). The classification of soil type uses the inputs listed in the last line of Table 2, although the available data only contains patterns of classes light and medium. A geographical plot of the soil type data is shown in the Figure 3 (right panel). Locations are widely distributed among the Jalna and Bid districts of the Marathwada region.

3.2. Machine learning techniques

The classifiers used in the current study, selected among the best ones in the comprehensive comparison (Fernández-Delgado et al., 2014), are implemented in Java using the Weka library (Hall et al., 2009), in R (R-Team, 2016)[4], in C++ and in Matlab (MathWorks, 2012)[5]. Henceforth, the suffix of the classifier name shows the implementation used: _w, _r, _c and _m mean Weka, R, C++ and Matlab, respectively. The classifiers, grouped by families, are described in the following enumeration alongside with their tunable metaparameters, whose values are listed in Table 3 (the symbol # means 'number of', e.g. #neurons means 'number of neurons in the hidden layer' of the ANNs).

1. *Bagging*: **bg_r**: bagging ensemble (Breiman, 1996) of decision tree base classifiers, implemented by the *ipred* (*bagging* function) and *rpart* R packages; **bgf_r**: bagging of flexible discriminant analysis (FDA) base classifiers implemented by the *earth* R package (*bagFDA* function), tuning the product degree and the #terms to prune; **bgp_w**: bagging of pruned par-

[4]https://www.r-project.org
[5]http://mathworks.com

Table 3: List of the tunable metaparameters and values of the classifiers in the form: number of values:start value - final value. The six functions of elm_m are sign, sine, sigmoid, hardlimit, radial basis and triangular basis.

Classifier	Metaparameters	Classifier	Metaparameters
bgf_r	degree=1,2 #terms=10:2 - 11	jrp_w	#folds=4:3 - 10, #runs=2,3,4
bgp_w	P=4:25 - 100	knn_r	#neighbors=13:1 - 37
dj48_w	#trees=3:10 - 20	mlp_m	#neurons=11:3 - 30
dtnb_w	#folds={1,4,5,10}	pnn_m	spread=19:0.01 - 10
elm_m	transf.funct=6:1 - 6,	rbf_m	spread=29:0.1 - 70
	#neurons=20:3 - 200	rf_w	#trees=4:100 - 750
gelm_m	C=20:$\{2^i\}_{-5}^{14}$,	rtf_w	#trees=5:10 - 50, perc=5:10 - 75
	spread=25:$\{2^i\}_{-16}^{8}$	svm_c	C=20:$\{2^i\}_{-5}^{14}$, spread=25:$\{2^i\}_{-16}^{8}$
j48_w	C=0.1,0.25,0.5		

tial C4.5 decision trees (PART) provided by Weka (Frank and Witten, 1998), tuning the bag size P.

2. *Boosting*: **ab_r**: Adaboost.M1 ensemble of classification trees (Alfaro et al., 2007), implemented by the *adabag* R package (*boosting* function).

3. *Decision trees*: **j48_w**: C4.5 decision tree (Quinlan, 1993) implemented in Weka, tuning the pruning confidence threshold C; **dj48_w**: decorate ensemble of J48 tree classifiers with high diversity (Melville and Mooney, 2004) provided by Weka, tuning the #trees; **rpt_r**: recursive partitioning (Breiman et al., 1984) implemented by the *rpart* R package (*rpart* function); **rt_w**: non-pruned random tree in Weka.

4. *Nearest neighbors classifiers*: **knn_r**: K-nearest neighbor (Ripley, 1996) implemented by the *class* R package (*knn* function), tuning the #neighbors.

5. *Neural networks*: **elm_m**: extreme learning machine (ELM), implemented by the publically available Matlab code[6], tuning the transfer function and the #neurons (Huang et al., 2012); **gelm_m**: ELM with Gaussian ker-

[6] http://extreme-learning-machines.org

nel tuning the regularization parameter C and the kernel spread; **mlp_m**: multi-layer perceptron neural network with the Matlab *train* function, tuning the #neurons; **pnn_m**: probabilistic neural network (Specht, 1990) with the *newpnn* Matlab function, tuning the Gaussian spread; **rbf_m**: radial basis function neural network (Park and Sandberg, 1993), implemented by the *newrb* Matlab function, tuning the Gaussian spread;

6. *Random forests*: **rf_r**: random forest ensemble (Breiman, 2001) of random tree base classifiers implemented by the *randomForest* R package; **rf_w**: RF implemented in Weka, tuning the #trees; **rtf_w**: rotation forest ensemble (Rodríguez and Kuncheva, 2006) of J48 base classifiers, tuning the #trees and the percentage of training patterns to be removed.
7. *Rule-based classifiers*: **dtnb_w**: decision table-naive Bayes hybrid classifier provided by Weka (Hall and Frank, 2008), tuning the #folds; **jrp_w**: repeated incremental pruning to produce error reduction (RIPPER), provided by Weka, tuning the #folds for reduced error pruning and the #runs for optimization (Cohen, 1995).
8. *Support vector machine*: **svm_c**: support vector classifier with Gaussian kernel implemented by LIBSVM[7] in C++, tuning the same parameters as gelm_m (Chang and Lin, 2011).

3.3. Experimental setup

The measure used for classification performance is the Cohen kappa (Viera and Garrett, 2005), henceforth denoted by κ and measured in %, which evaluates the classification accuracy discarding the probability of classifier success by chance, which is defined by:

[7] http://www.csie.ntu.edu.tw/∼cjlin/libsvm

$$\kappa = 100\frac{p_a - p_e}{s - p_e}$$

$$p_a = \sum_{i=1}^{C} n_{ii}; \quad p_e = \frac{1}{s}\sum_{i=1}^{C}\left(\sum_{j=1}^{C} n_{ij}\right)\left(\sum_{k=1}^{C} n_{ki}\right); \quad s = \sum_{i=1}^{C}\sum_{j=1}^{C} n_{ij}$$

where C is the number of classes and n_{ij} is the number of patterns of class i assigned by the classifier to class j, for $i, j = 1, \ldots, C$. We also use two additional class-specific performance measures: sensitivity (SE) and positive predictivity (PP) of each class i, defined as $SE_i = 100 n_{ii} / \sum_{j=1}^{C} n_{ij}$ (percentage of patterns of class i assigned by the classifier to class i) and $PP_i = 100 n_{ii} / \sum_{j=1}^{C} n_{ji}$ (percentage of patterns assigned by the classifier to class i which really belong to class i). For each classification problem, we developed a 4-fold validation creating four groups of data sets, each composed by a training, a validation and a test set which do not intersect among them. The patterns of each class are randomly shuffled and 50% of them are used for training, 25% for validation and 25% for testing, thus guaranteeing that the same percentages of the pattern set are devoted to training, validation and test sets. For each class, the three sets are rotated among folds: e.g. with M patterns, the first fold uses patterns from 1st to $M/2$-th for training, patterns from $M/2 + 1$ to $3M/4$ for validation, and patterns from $3M/4 + 1$ to M for testing; the second fold uses patterns from $M/4+1$ to $3M/4$ for training, from $3M/4+1$ to M for validation and from 1 to $M/4$ for testing; and so on. Each classifier is trained on the four training sets using each combination of values of its metaparameters (Table 3), and then it is tested on the validation sets. For each combination of metaparameter values, the average κ over the four validation sets is calculated, and the combination of metaparameter values with the best κ is selected for testing. At this point, each classifier is trained on each training set with the selected metaparameter values, and then it classifies each test set. The final test performance is the average κ over the four test sets. Note that any pattern of the test set is not included in the training nor in the validation sets of the same group, so the test results can not be optimisticaly biased by overfitting.

4. Results and discussion

Table 2 lists the information (number of patterns and inputs) of each classification problem, and the population of each class. Given that the region of Marathwada has soil with low OC-F and Fe-F levels, the low class is the most populated for both problems. This soil exhibits medium P_2O_5 level predominantly and it is slightly alkaline, so the SAL class is the most populated. Besides, most of the soil is of type medium, and the major crops are soybean(R) and cotton(R), which rely on rainfall because Marathwada comes under drought area since four years. The data are preprocessed in order to have zero mean and one standard deviation for each input.

Table 4: Values of κ (in %) achieved by each classifier for each classification problem. The best κ for each problem is in bold.

Classifier	OC-F	P_2O_5-F	Mn-F	Fe-F	N_2O	P_2O_5	K_2O	pH	Crop	Soil
ab_r	88.50	**85.54**	59.33	67.45	25.40	**35.08**	25.81	44.71	85.27	96.65
bg_r	73.84	74.95	52.72	53.58	28.63	27.42	28.63	26.76	82.02	93.47
bgf_r	85.18	78.77	54.95	64.21	26.61	29.03	29.44	40.06	77.79	95.47
bgp_w	82.42	74.97	54.95	64.09	24.60	24.60	20.97	42.13	86.37	95.49
dj48_w	83.03	80.45	49.44	57.18	31.85	28.63	24.60	42.49	85.15	95.37
dtnb_w	81.33	66.78	47.79	44.69	22.18	22.58	5.24	35.93	85.85	**97.82**
elm_m	75.45	68.99	48.41	49.90	23.39	25.81	12.10	34.61	82.25	91.76
gelm_m	82.41	77.23	57.14	57.85	30.65	30.24	20.16	42.91	85.17	96.20
j48_w	69.25	70.76	37.92	54.42	20.16	21.37	14.92	40.78	82.34	94.76
jrp_w	71.28	70.34	45.02	52.00	19.35	22.18	12.90	40.44	83.12	94.94
knn_r	74.14	72.56	52.77	54.09	25.81	24.19	15.73	41.55	84.32	94.46
mlp_m	56.95	30.25	38.77	41.46	24.60	17.74	12.10	13.73	47.62	86.18
pnn_m	77.43	73.33	55.56	52.82	23.79	22.98	18.15	40.38	84.52	94.90
rbf_m	48.82	40.04	34.40	43.32	2.82	14.92	2.82	13.30	82.99	92.03
rf_r	87.35	83.01	58.78	**69.35**	**33.06**	33.06	**31.85**	**47.32**	**88.13**	96.37
rf_w	**90.65**	79.58	64.27	65.17	30.24	32.26	26.61	46.85	87.64	96.80
rpt_r	67.05	63.81	46.65	36.23	21.37	22.58	21.37	38.92	80.01	92.89
rt_r	69.15	68.93	51.14	36.03	15.32	21.37	19.35	33.49	78.75	93.39
rtf_w	87.39	75.83	57.70	59.01	28.23	30.65	27.42	46.77	86.34	96.80
svm_c	80.2	82.3	**64.8**	60.1	29.0	30.2	18.5	43.0	86.1	95.8

Table 5: Columns 1-6: classifier ranking and position according to the Friedman rank test. Columns 7-12: classifiers ordered by decreasing p-value of a Wilcoxon signed rank test comparing the best ranked classifier (rf_r) to the remaining ones (the significant tests are in bold). The label — means that rf_r can not be compared to itself.

Friedman rank test						Wilcoxon signed rank test					
Pos.	Clasif.	Rank	Pos.	Clasif.	Rank	Pos.	Clasif	p-value	Pos.	Clasif	p-value
1	rf_r	2.200	11	knn_r	11.400	1	—	—	11	dtnb_w	0.384494
2	rf_w	2.950	12	pnn_m	11.400	2	ab_r	0.969839	12	jrp_w	0.384494
3	ab_r	3.900	13	dtnb_w	12.650	3	rf_w	0.850051	13	elm_m	0.344523
4	rtf_w	4.650	14	jrp_w	14.300	4	rtf_w	0.623046	14	j48_w	0.307308
5	svm_c	5.900	15	j48_w	14.350	5	svm_c	0.520366	15	elm_m	0.34055
6	gelm_m	6.600	16	elm_m	14.650	6	dj48_w	0.495968	16	rt_w	0.272856
7	dj48_w	7.400	17	rpt_r	15.650	7	bgf_r	0.472509	17	bg_r	0.272675
8	bgf_r	8.050	18	rt_w	16.250	8	bgp_w	0.472342	18	rpt_r	0.240968
9	bgp_w	9.600	19	mlp_m	18.200	9	knn_r	0.427181	19	rbf_m	0.10385
10	bg_r	11.300	20	rbf_m	18.600	10	pnn_m	0.427181	20	mlp_m	**0.03114**

4.1. Global discussion of the results

Table 4 reports the κ achieved by each classifier for the 10 classification problems: OC-F, P_2O_5-F, Mn-F and Fe-F (village-wise soil fertility indices of nutrients OC, P_2O_5, Mn and Fe); nutrients N_2O, P_2O_5 and K_2O; soil pH; suitable crop and soil type. Globally, rf_r and rf_w achieve the best κ for 5 and 1 problems, respectively, being very near to the best ones for the other 4 problems. The ab_r is the best in two problems (P_2O_5-F and P_2O_5), while svm_c and dtnb_w are the best for one problem each (OC-F, Mn-F and soil, respectively). The columns 1-6 of Table 5 reports the classifiers ordered by their Friedman rank test (decreasing with the classifier performance): the rf_r (Random Forest in R) is the best with a rank of 2.2, which means that in average it is nearly the second best classifier for all the problems. The RF provided by Weka (rf_w) is the second one, so its metaparameter tuning does not improve the good results of the R version. Adaboost in R (ab_r) and Rotation Forest of J48

trees in Weka (rtf_w) also achieve good ranks, although far from rf_r (1.7 and 2.4 points higher), and svm_c gets the 5th position (3.7 points higher), followed by gelm_m (Gaussian kernel Extreme Learning Machine in Matlab, 6th position), which works much better than elm_m (16th position). The bagging classifiers bgf_r, bgp_w and bg_r work similarly (positions 8, 9 and 10, respectively), and similarly to knn_r (K-nearest neighbor classifier in R) and pnn_m (probabilistic neural network in Matlab). Considering columns 7-12, the Wilcoxon signed rank tests between rf_r and the other classifiers show that $p > 0.05$ (i.e., the differences are not statistically significant) except with respect to mlp_m, which gets the worst results. The lack of statistical significance might be due to the low number of measurements (just ten classification problems) for each classifier.

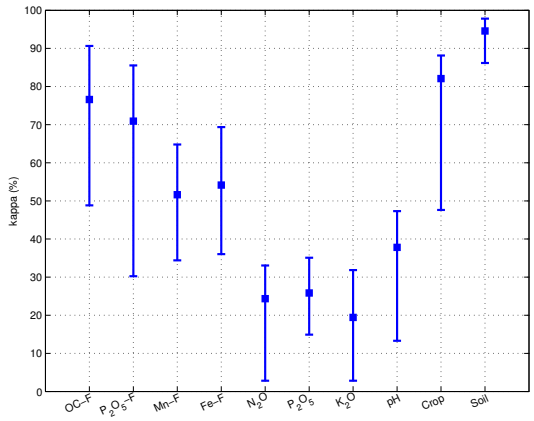

Figure 4: Intervals of κ (in %) of the different classifiers for each problem (the filled square shows the mean κ).

Figure 4 shows the κ intervals (minimum, mean and maximum) achieved by the different classifiers for each problem, with large differences in the κ values and in the interval width. The soil type classification problem has the highest maximum κ (upper limit of the interval) and the narrowest interval (i.e., all the classifiers work very well), while OC-F, P_2O_5-F and crop have high maximum κ with wider intervals. The N_2O, P_2O_5, K_2O and pH have lower κ values.

The intervals for each classifier over all the problems are reported by Figure

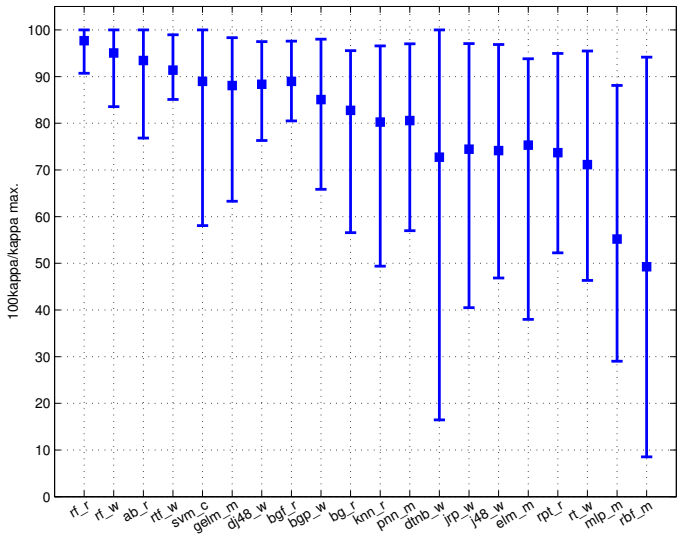

Figure 5: Intervals of percentages of the maximum κ for each problem, for all the problems and for each classifier.

5 but, since the κ values are very different among problems, each interval represents the percentages of the maximums κ in the different problems achieved by the classifier. For example, the maximum κ in the OC-F problem is 90.65% achieved by rf_w, so rf_w has 100% of the maximum κ, while ab_r has 100 $\cdot$ 88.5/90.65=97.62% of the maximum κ. In Figure 5, the rf_r has the highest and narrowest interval, being above 90% of the maximum κ for all the problems. Some other classifiers as ab_r, svm_c or dtnb_w have intervals whose maximum achieves 100%, but their means and minimums are much lower than rf_r (e.g. dtnb_r achieves about 15% of the maximum κ for some problem). Although almost all the classifiers achieve more than 90% of the maximum κ for some data set (excepting mlp_m), all the classifiers exhibit minimum percentages below 80% excepting rf_r, rf_w, rtf_w and bgf_r.

4.2. Classification of village-wise soil OC, P_2O_5, Mn and Fe fertility indices

The confusion matrices (in %) of the best classifiers, along with κ, SE and PP are reported by Table 6 for the soil fertility indices problems OC-F, P_2O_5-

Table 6: Confusion matrices (in %) of the best classifiers (showed in the matrix header between brackets) for the classification of soil fertility indices. The symbol "—" means that the class is not available. Labels L, M and H mean low, medium and high respectively.

	Major nutrients					Micro-nutrients			
	OC-F(rf_w)		P_2O_5-F(ab_r)			Mn-F(svm_c)		Fe-F(rf_r)	
Fertility	L	M	L	M	H	M	H	L	M
L	**52.9**	1.4	**22.8**	3	0.3	—	—	**66.2**	2.8
M	3.3	**42.4**	1.3	**58.5**	0.3	**36.5**	8.5	7.5	**16.5**
H	—	—	0.3	1.5	**5.3**	7.5	**38.5**	—	—
κ (%)	90.65		85.54			64.8		69.35	
SE(%)	97.5	92.9	87.5	97.5	75	81.1	83.7	96	68.8
PP(%)	94.2	96.9	93.8	92.9	91.3	83	81.9	89.8	85.7

F, Mn-F and Fe-F. The rf_w achieves the best κ for OC-F (90.65%, Table 4), followed by ab_r (88.5%), rtf_w (87.39%) and rf_r (87.35%). The confusion matrix of rf_w shows low values outside the diagonal, being the percentage of M patterns assigned to class L (3.3%) twice the percentage of L patterns assigned to class M (1.4%). The classes L and M exhibit high SE and PP. For P_2O_5-F, the best classifier is ab_r (κ=85.54%), followed at certain distance by rf_r (83.01%) and svm_c (82.3%). In this case, the errors in the confusion matrix of ab_r are for adjacent classes, e.g., between L and M, or between M and H, but the error percentages between non-adjacent classes are low (0.3%). The ab_r tends to classify L patterns as M (3%), and in less degree H patterns as M (1.5%) and M patterns as L (1.3%). The best detected class is M, with SE and PP above 93%, while SE of class L (87.5%) and, specially, of class H (75%) are much lower. The results are worse for the classification of Mn-F and Fe-F problems. For Mn-F, the svm_c is the best classifier (κ=64.8%), followed by rf_w (64.27%), and the following classifiers are under 60% (ab_r achieves 59.33%). The confusion matrix of svm_c shows diagonal values (above 36%) much higher than outside the diagonal (7-8%), with slightly higher percentage of M patterns classified as H than the opposite. The results for Fe-F are slightly better: rf_r achieves

κ=69.35%, followed by ab_r (67.45%) and rf_w (65.17%). The confusion matrix of rf_r shows a high number of M patterns classified as L (7.5%), while class M, which is much less populated, exhibits a low SE (68.8%).

Table 7: Confusion matrices (in %) of the best classifiers for N_2O, P_2O_5 and K_2O. Labels L, M and H mean low, medium and high, respectively. The values in the matrix diagonal are in bold.

	N_2O(rf_r)			P_2O_5(ab_r)			K_2O(rf_r)		
	L	M	H	L	M	H	L	M	H
L	**21.77**	8.33	3.23	**22.04**	6.99	4.30	**19.35**	9.41	4.57
M	11.02	**13.17**	9.14	9.41	**14.78**	9.14	8.87	**14.25**	10.22
H	6.45	6.45	**20.43**	4.57	8.87	**19.89**	4.57	7.80	**20.97**
κ(%)		33.06			35.08			31.83	
Acc(%)		55.37			56.71			54.56	
SE(%)	65.3	39.5	61.3	66.1	44.4	59.7	58.1	42.7	62.9
PP(%)	55.5	47.1	62.3	61.2	48.2	59.7	59.0	45.3	58.6

4.3. Classification of soil nutrients N_2O, P_2O_5 and K_2O

Table 7 reports the best confusion matrices for the classification of N_2O and K_2O, achieved by rf_r (κ=33.06% and 31.83%, respectively), and for the classification of P_2O_5, achieved by ab_r (κ=38.08%), results which are worse than the corresponding to the classification of the village-wise soil fertility indices. For the three nutrients the class M has the lowest SE and PP (between 39-48%), while classes L and H exhibit SE and PP values about 60%. Since the classes represent quantization levels, they can be considered as ordinal classification problems, with an ordering relation among the classes. This explains that classes L and H are better recognized, while the middle class M wins external or loses internal patterns from/to L and H. Besides, the three matrices exhibit the highest non-diagonal values in positions adjacent to the diagonal, because the classification errors are more probable between adjacent classes representing contiguous levels of quantization.

4.4. Soil pH classification

The rf_r achieves the best performance for this problem, with κ=47.32% (Table 4) and the confusion matrix of Table 8, for which the accuracy is 69.63%. Again, the only high values outside the diagonal correspond to adjacent classes: SA and N, with values 0.64% and 0.75% where the diagonal value is 0.37%; N and SAL, being more probable to classify N patterns as SAL (10.45%) than the opposite (6.66%); SAL and MAL, with more MAL patterns (7.78%) classified as SAL than MAL patterns correctly classified (2.67%). The remaining non-diagonal values are below 1%. The SE and PP are only acceptable (above 68%) for the most populated classes N and SAL, while the less populated classes SA and MAL exhibit worse results (SE about 20%).

Table 8: Confusion matrix (in %) of the best classifier (rf_r, with κ= 47.32% and accuracy= 69.63%) for soil pH classification. Labels SA, N, SAL and MAL mean slightly acidic, neutral, slightly alkaline and moderately alkaline respectively.

pH	SA	N	SAL	MAL	SE(%)	PP(%)
SA	**0.37**	0.64	0.27	0	29.2	26.9
N	0.75	**26.92**	10.45	0.27	70.1	74.6
SAL	0.21	6.66	**39.66**	1.44	82.7	68.2
MAL	0.05	1.87	7.78	**2.67**	21.6	61.0

4.5. Crop classification

The rf_r achieves the best κ (88.13%) also for crop classification (Table 4), followed by rf_w (87.64%), bgp_w (86.37%) and rtf_w (86.34%). The ab_r and svm_c, which achieved good results in the previous problems, also work well. The confusion matrix of the rf_r for this problem (Table 9) has almost all the values outside the diagonal under 1%. Only the percentage (3.34%) of bajra(R) patterns classified as cotton(R) is higher than the corresponding diagonal term (2.89%), which reduces the SE of class bajra(R) to 45.1%. The reason is that the usual crop rotation in Marathwada is between both classes, so their patterns are very similar. Besides, cotton(R) is more important and populated than

bajra(R), whose patterns tend to be classified as cotton(R). All the remaining classes have SE above 90% and PP above 80%.

Table 9: Confusion matrix (in %) of rf_r for crop classification (κ=88.13%, accuracy=93.09%). Labels R and I mean rainfed and irrigated respectively.

Crop	Bajra(R)	Cotton(I)	Cotton(R)	Soybean(R)	SE(%)	PP(%)
Bajra(R)	**2.89**	0.04	3.34	0.14	45.1	79.8
Cotton(I)	0	**10.14**	0.35	0.76	90.1	95.7
Cotton(R)	0.74	0.07	**23.14**	0.81	93.5	85.3
Soybean(R)	0	0.34	0.32	**56.90**	97.1	97.1

4.6. Soil type classification

The best soil type classifier is dtnb_w (hybrid classifier of decision table and naive Bayes), with κ=97.82%, followed by the group of methods which were the best in the previous classification problems: rf_w and rtf_w (96.80%), ab_r (96.65%), rf_r (96.37%) and svm_c (95.8%). The confusion matrix of the two best classifiers (Table 10) shows similarly low non-diagonal values for classes L and M, with SE and PP values above 97%.

Table 10: The confusion matrix (in %) of the two best classifiers (dtnb_w and rf_w, κ= 97.82% and 96.80% respectively) for soil type classification. Labels L and M mean light and medium respectively.

	dtnb_w				rf_w			
Soil	L	M	SE(%)	PP(%)	L	M	SE(%)	PP(%)
L	**27.96**	0.47	98.33	99.42	**27.73**	0.71	97.50	97.17
M	0.41	**71.15**	98.54	99.34	0.59	**70.97**	97.91	99.01

4.7. Classification comparison among regions

As well as the data from the region of Marathwada, we also have data availabe from regions Paschim-Maharashtra and North-Maharashtra, in the same

state of Maharashtra. An interesting issue is how valid are the classifiers, trained using data from one region, to test data from different regions. In other words, what is the representativity and generalization ability of the trained classifiers with respect to regions? We developed experiments training and tuning the metaparameters of the classifiers with patterns from one region, and then testing the trained and tuned classifier with data from different regions. Specifically, we have data from Marathwada, North Maharashtra and Paschim Maharashtra (henceforth labeled as M, NM and PM, respectively) highlighted in Figure 2, for village-wise OC-F, P_2O_5-F, Mn-F, Fe-F and pH problems. The corresponding experiments for N_2O, P_2O_5 and K_2O, crop and soil type classification could not be developed due to the lack of data for regions NM and PM.

Table 11: Values of κ (in %) achieved by the best classifier training and testing with patterns of different regions (first column, see text for region labels). The symbol '—' means that data are not available.

	OC-F		P_2O_5-F		Mn-F		Fe-F		pH	
Regions	Best	κ	Best	κ	Best	κ	Best	κ	Best	κ
M-NM	rpt_r	17.21	bgp_r	15	bg_r	65.98	rpt_r	66.78	rtf_w	23.97
M-PM	elm_m	9.66	rf_r	100	bg_r	66.62	rpt_r	70.52	rtf_w	66.22
NM-M	rtf_w	32.60	rf_r	34.10	rt_w	45.70	svm_c	42.50	—	—
NM-PM	pnn_m	14.99	bg_r	100	bg_r	100	j48_w	100	—	—
PM-M	rt_w	12.73	bgf_r	34.57	bgf_r	45.15	elm_m	44.33	knn_r	69.61
PM-NM	pnn_m	11.13	bg_r	100	bg_r	100	svm_c	100	rf_r	48.23

Table 11 reports the best classifier for each classification problem and combination of training and test regions, and the κ that it achieves (e.g. M-PM in the leftmost column means training and test using data from Marathwada and Paschim-Maharashtra respectively). No classifier achieves good results for OC-F, no matter the combination of regions used for training and testing, and the best result is κ=32.6% with rtf_w. The results for P_2O_5-F are very good (κ=100%) for NM-PM and PM-NM (both with bg_r) and for M-PM (with rf_r), but much worse (between 15% and 35%) for the remaining combinations. This is surprising, because M-PM works well, but PM-M works much worse, sug-

gesting that data from region M are somehow representative for data from PM, but the opposite is not true. For Mn-F and Fe-F problems, both NM-PM and PM-NM work well (100% with bg_r for Mn-F, and with j48_w and svm_c for Fe-F), so the data seem to be valid between NM and PM. The performance of cases M-NM and M-PM are about 66-71%, so the data from M are somehow valid for the other two regions, but the inverse combinations (NM-M and PM-M) are about 42-46%, so NM and PM data are not so valid for region M. As a conclusion: 1) the data for OC-F classification are not valid among regions; 2) the data for P_2O_5-F, Mn-F and Fe-F classifications are valid only between North Maharashtra and Paschim Maharashtra, and from Maharashtra to Paschim-Maharashtra; 3) the data for P_2O_5-F classification are compatible from Marathwada to Paschim Maharashtra, but not the inverse; and 4) the data for Mn-F and Fe-F classifications are slightly compatible (about 66-70%) from Marathwada to the other regions, but not the inverse.

5. Conclusions

Agriculture is a major sector in the Indian economy, which is affected by changing trends in temperature and rainfall, insufficient water, agriculture practices and nutrient deficiences. Adequate soil parameters and proper application of fertilizers may help to attenuate these problems. The current research supports the Indian Government to make decisions about improving soil quality and crop production. The soil quality depends on its type and pH, village-wise fertility indices of OC, P_2O_5, Mn and Fe, and on the selected crop. Thus, the classification of their values from measurements of N_2O, P_2O_5, K_2O, SO_4 and EC, among others, allows to reduce the cost of the chemical analysis and to save time for specialized technicians. The values of soil nutrients N_2O, P_2O_5 and K_2O are also very useful for recommendation of fertilizers. Thus, a wide collection of diverse classifiers is used to predict their levels (low, medium and high) using the previous inputs. This collection includes bagging and boosting ensembles, random forests, neural networks, support vector machine, decision

trees, rule-based and nearest neighbor classifiers. We achieved values of the Cohen κ parameter above 90% for the village-wise OC fertility index and soil classification, and above 85% for P_2O_5 fertility index and crop classification, while the classification of Mn and Fe fertility indices achieved nearly 65%. The κ is about 47% for pH classification, and 35% for N_2O, P_2O_5 and K_2O classification, with accuracies about 55%. The R-version of the random forest classifier is the first in a Friedman rank test, being the best in 5 of 10 problems and overcoming 90% of the maximum accuracy in all the cases, although the difference with almost all the remaining classifiers is not statistically significant. Other classifiers with good results are the Weka version of random forest, the R version of adaboost, the Weka rotation forest of J48 base classifiers, the LIBSVM support vector machine and the extreme learning machine, both with Gaussian kernels. The remaining methods work much worse. We studied the model validity across three Indian regions (Marathwada, North Maharashtra and Paschim Maharashtra), training and testing each classifier with data from different regions, finding compatibilities between North Maharashtra and Paschim Maharashtra for village-wise P_2O_5, Mn and Fe fertility indices classification. The results of this study might contribute to design agricultural strategies of the Indian Government to manage the soil fertility degradation, crop productivity and usage of fertilizers. Future work includes to use these methods to create village-wise soil fertility maps for OC, P_2O_5, Mn and Fe nutrients.

Funding

This research is supported by the Erasmus Mundus Euphrates programme [project number 2013-2540/001-001-EMA2].

References

Alfaro, E., Gámez, M., García, N., 2007. Multiclass corporate failure prediction by Adaboost.M1. Int. Advances in Economic Res. 13, 301–312.

Bowman, R., Guenzi, W., Savory, D., 1991. Spectroscopic method for estimation of soil organic carbon. Soil Sci. Soc. of Am. J. 55, 563–566.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees.

Brungard, C., Boettinger, J., Duniway, M., Wills, S., Jr., T.E., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239240, 68–83.

Chang, C., Lin, C., 2011. LIBSVM: a library for support vector machines. ACM Trans. on Intel. Syst. and Technol. 2, 27:1–27:27.

Chesworth, W., 2008. Encyclopedia of Soil Science. Springer.

Chtiouia, Y., Panigrahia, S., Francl, L., 1999. A generalized regression neural network and its application for leaf wetness prediction to forecast plant disease. Chem. and Intel. Lab. Systs. 48, 47–58.

CIA, 2016. The world factbook 2016-2017. Central Intelligence Agency, Washington D.C.

Cohen, W., 1995. Fast effective rule induction, in: Int. Conf. on Mach. Learn., pp. 115–123.

Department of Agriculture & Cooperation, 2011. Methods manual. Soil testing in India. Technical Report. Ministry of Agriculture, Government of India. New Delhi.

Directorate of Economics and Statistics, 2015. Economic survey of Maharashtra. Technical Report. Planning Department, Maharashtra Government. Mumbai.

Feller, C., Blanchart, E., Bernoux, M., Lal, R., Manlay, R., 2012. Soil fertility concepts over the past two centuries: the importance attributed to soil organic

matter in developed and developing countries. Archives of Agronomy and Soil Science 58(S1), S3–S21.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real classification problems? J. Mach. Learn. Res. 15, 3133–3181.

Ford, C., 1954. Determination of sodium and potassium oxides by flame photometry. Anal. chem. 26, 1578–1581.

Frank, E., Witten, I., 1998. Generating accurate rule sets without global optimization, in: Int. Conf. on Mach. Learn., pp. 144–151.

Groenigen, J., Huygens, D., Boeckx, P., Kuyper, T., Lubbers, I., Rtting, T., Groffman, P., 2015. The soil N cycle: new insights and key challenges. Soil 1, 235–256.

Guerrero, J., Pajares, G., Montalvo, M., Romeo, J., Guijarro, M., 2012. Support vector machines for crop/weeds identification in maize fields. Expert Syst. Appl. 39, 11149–11155.

Hall, M., Frank, E., 2008. Combining naive Bayes and decision tables. Proc. Artif. Intel. Soc. Conf. , 318–319.

Hall, M., Frank, E., G.Ho., Pfahringer, B., Reutemann, P., Witten, I., 2009. The Weka data mining software: an update. SIGKDD Explorations 11, 10–18.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C., Schmidt, M., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77.

Hill, M., Connolly, P., Reutemann, P., Fletcher, D., 2014. The use of data mining to assist crop protection decisions on kiwifruit in New Zealand. Comput. Electron. Agric. 108, 250–257.

Huang, G.B., Zhou, H., Ding, X., Zhang, R., 2012. Extreme learning machine for regression and multiclass classification. IEEE Trans. Systs., Man, and Cybern.-Part B: Cybern. 42(2), 513–529.

Jackson, M., 1958. Soil chemical analysis. Prentice Hall of India Pvt. Ltd., New Delhi.

Jakubauskas, M., Legates, D., Kastens, J., 2002. Crop identification using harmonic analysis of time-series AVHRR NDVI data. Comput. Electron. Agric. 37, 127–139.

Jones, C., 1912. Activity of organic nitrogen as measured by the alkaline permanganate method. J. Ind. Eng. Chem. 24, 438–441.

Katyal, J., Rattan, R., 2003. Secondary and micronutrients: research gaps and future needs. Fertil. News 48, 9–20.

Leggeit, Argyle, D., 1983. The DTPA-extractable iron, manganese, copper, and zinc from neutral and calcareous soils dried under different conditions. Soil Sci. Soc. Am. J. 47(3), 518–522.

Mahatma Phule Agricultural University, 2016. Krishi Darshani. Parbhani, Maharashtra, India. Page 18 (in Marathi language).

MathWorks, 2012. version 7.14 (R2012a). Natick, Massachusetts.

Melville, P., Mooney, R.J., 2004. Creating diversity in ensembles using artificial data. Inform. Fusion: special issue on diversity in multiclassifier systems 6, 99–111.

Mkhabela, M., Bullock, P., Raj, S., Wang, S., Yang, Y., 2011. Crop yield forecasting on the Canadian prairies using MODIS NDVI data. Agric. For. Meteorol. 151, 385–393.

Mucherino, A., Papajorgji, P., Pardalos, P., 2009. A survey of data mining techniques applied to agriculture. Oper. Res. 9, 121–140.

Muhr, G., Datta, N., Shankara, S., Dever, F., Lecy, V., Donahue, R., 1965. Soil testing in India. U.S. Agency for International Development, Mission to India.

Naidu, L., Ramamuthy, V., Sidhu, G., Sarkar, D., 2011. Emerging deficiency of potassium in soils and crops of india. Karnataka J. Agric. Sci. 24, 12–19.

Narkhede, U., Adhiya, K., 2014. A study of clustering techniques for crop prediction - a survey. Am. Int. J. of Res. in Sci., Technol., Eng. and Math. 5(1), 44–48.

Obade, V., Lal, R., 2016. Towards a standard technique for soil quality assessment. Geoderma 265, 96–102.

Olsen, S., 1954. Estimation of available phosphorus in soils by extraction with sodium bicarbonate. Circular series, U.S. Dept. of Agriculture.

Panigrahy, S., Sharma, S., 1997. Mapping of crop rotation using multidate Indian remote sensing satellite digital data. ISPRS J. of Photogrammetry and Remote Sens. 52, 85–91.

Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., Mouazen, A., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121, 57–65.

Park, J., Sandberg, I., 1993. Approximation and radial-basis-function networks. Neural Computation 5(2), 305–316.

Pea, M., Brenning, A., 2015. Assessing fruit-tree crop classification from Landsat-8 time series for the Maipo Valley, Chile. Remote Sens. Environ. 171, 234–244.

Quinlan, R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

R-Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, version 3.1.2. Vienna, Austria.

Rammoorthy, B., Bajaj, J., 1969. Available N, P and K status of Indian soils. Fertilizer news 14(8), 24–26.

Rashidi, M., Seilsepour, M., 2009. Modeling of soil total nitrogen based on soil organic carbon. ARPN J. of Agric. and Biol. Sci. 4(2), 1–5.

Reeves, D., 1997. The role of soil organic matter in maintaining soil quality in continuous cropping systems. Mach. Learn. 43, 131–167.

Richards, L., Allison, L., Bernstein, L., Bower, C., Brown, J., Fireman, M., Hatcher, J., Hayward, H., Pearson, G., Reeve, R., Richards, A., Wilcox, L., 1954. Diagnosis and improvement of saline and alkaline soils. Science 12, 800.

Ripley, B., 1996. Pattern Recognition and Neural Networks. Cambridge Univ. Press.

Rodríguez, J., Kuncheva, L., 2006. Rotation forest: A new classifier ensemble method. IEEE Trans. on Pat. Anal. and Mach. Intel. 28(10), 1619–1630.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., Roberts, D., 2008. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. Remote Sens. Environ. 112, 2272–2283.

Romero, J., Roncallo, P., Akkiraju, P., Ponzoni, I., Echenique, V., Carballido, J., 2013. Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. Comput. Electron. Agric. 96, 173–179.

Sehgal, J., 1990. Agro-ecological Regions of India. Technical bulletin (National Bureau of Soil Survey & Land Use Planning), Indian Council of Agricultural Research.

Sheela, P., Sivaranjani, K., 2015. A brief survey of classification techniques applied to soil fertility prediction, in: Int. Conf. Eng. Trends in Sci. and Hum., pp. 80–83.

Specht, D., 1990. Probabilistic neural networks. Neural Netw. 3, 109–118.

Subbaiah, B., Asija, G., 1956. A rapid procedure for the estimation of available nitrogen in soil. Current Sci. 25, 259–260.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B., Triantafilis, J., 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma 253–254, 67–77.

Tatsumi, K., Yamashiki, Y., Torres, M.C., Taipe, C.R., 2015. Crop classification of upland fields using random forest of time-series Landsat 7 ETM+ data. Comput. Electron. Agric. 115, 171–179.

Turmel, M.S., Speratti, A., Baudron, F., Verhulst, N., Govaerts, B., 2015. Crop residue management and soil health: A systems analysis. Agric. Systs 134, 6–16.

Viera, A., Garrett, J., 2005. Understanding interobserver agreement: the kappa statistic. Family Med. 37(5), 360–363.