Method for dataset preparation for soil data analysis in decision support applications

Sanjay Motia¹, SRN Reddy²

¹USICT, Guru Gobind Singh Indraprastha University, Dwarka, Delhi, India
²Deptt. of CSE, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi, India

motia.sanjay@gmail.com
rammallik@yahoo.com

Abstract. Nowadays, publically available soil health cards (SHC) are easily understandable and useful means of knowledge transfer about soil conditions to farmers for sustainable soil management in farms. The SHCs are the true representative of precise and useful information about the physicochemical properties and nutrient status of agricultural soil. Before SHCs, the soil test records generally lie with soil test laboratories (STLs) and their availability in form of soil datasets is very limited. Moreover, the datasets are not big enough to be utilized as test datasets in soil investigation and development of futuristic and intelligent decision support and advisory systems. To address the issue, this paper presents a simple method for soil dataset preparation from SHCs for machine learning (ML) and artificial intelligence (AI) based soil research and analysis applications. The method collects wide number of SHCs for particular location and applies simple parsing and data cleaning concepts, implemented by utilizing open-source python libraries, to set of SHCs to prepare soil dataset. The dataset prepared by the proposed method can be used in a variety of intelligent data-driven decision support applications such as soil classification, nutrient prediction, and fertilizer recommendation. The dataset thus generated by the proposed method would help both researchers and soil scientists in their future research on soil health monitoring and decision support for sustainable farm management.

Keywords. Dataset Preparation, XML Parsing, Data Cleaning, Soil Health Card, Soil Health Analysis.

1. Introduction
The characterization of soil in terms of its fertility status is an important aspect of the sustainable growth of the agriculture production system. In the majority of agriculture dominated countries, soil nutrient status or soil health is generally assessed through laboratory analysis conducted on soil samples [1]. Post analysis, the soil physio-chemical properties are evaluated according to predefined soil fertility standards, and a soil health report is generated. In many developing countries, this report card of agricultural soil of any farm is termed as soil health card (SHC) [2]. Before the launch of the SHC scheme [3] by the Govt. of India, the results of soil testing usually lie with the respective STL and generally not available in the public domain. The SHC consists of detailed information on soil structure, physicochemical properties, nutrient status, and fertilizer prescription. Since the SHC is an outcome of soil testing, therefore its content is similar to the conventional laboratory-based soil test reports for a particular field.
Nowadays, the SHCs are publicly available and are serving as an instrumental aid in selecting area specific crops, judicious use of nutrients, increasing yield and improving the land use pattern for crop diversification, and sustaining soil fertility [4][5]. The information in SHC can’t be directly used for research purposes in the soft computing domain. It needs to be processed further before its usage in computing experiments.

In the present as well as futuristic artificial intelligence-based agriculture applications a huge amount of data would be required for analysis and intelligent decision making. The unavailability of site-specific training soil dataset is one of the main challenges for data analysts as well as researchers’. To help in addressing this issue, the paper proposes a method for dataset preparation from SHC for ML and AI based soil research and analysis applications [6]. The method reads hundreds or thousands of SHCs for particular area covering village/s, districts and applies simple XML parsing and data cleaning concepts, implemented by utilizing open-source python libraries, to set of SHCs to prepare soil dataset. The dataset thus prepared can be utilized in a wide variety of intelligent soil analysis based data analytic applications such as soil quality assessment [11], classification, nutrient prediction, and fertilizer recommendation. The dataset thus generated by the proposed method would help both researchers and soil scientists in their future research on soil health monitoring, evaluation and decision support for sustainable farm management.

The remaining part of the paper is categorized into sections where each section discussed the various aspects of proposed method. The section-2 gives insight into the problem addressed followed by brief discussion on SHC and its applications in section-3. The proposed methodology comprising of steps involved, implementation is explained in section-4. The results of experimentation are presented in Section-5 whereas the conclusion and future scope is given in section-6.

2. Problem Addressed

Data collection and pre-processing is a primary task to provide efficient and sufficient data for any data analytics application based on machine and deep learning models. A good amount of training dataset is required for model building and more accurate results. The existing soil datasets are somewhat more related to foreign subcontinent and are different in soil ecology as far as Indian soils condition are concerned. Furthermore, most of the soil test results lie with STLs, and availability of soil datasets having complete and detailed information on soil properties and nutrient contents is very limited. Also, the available Indian soil data sets are not huge enough to be utilized in AI and ML-based data analytic applications. The availability of large datasets is pre-requisite for accurate and good analysis results for any decision support cum recommendation application. Therefore, to implement an ML model for intelligent fertigation advisory system for Indian soils, the prerequisite requirement is the availability of larger dataset that is rich in soil physicochemical and nutrient specifications and obtained as a result of the application of method covering:

a) Collection of village-wise, district or location-specific SHCs in available data format,
b) Soft-method that reads and transforms (if required) the available SHC into desired data format,
c) Supports reading and processing of multiple numbers (hundreds or thousands) of SHCs,
d) Process all the data contained in the SHCs (transformation, cleaning, format conversion) and finally
e) Formation and presentation of the dataset. The data can be used by a data analyst to perform a detailed-level of analysis for fertilizer recommendation.

To help in meeting the above-mentioned goals, this paper proposes a method to develop soil data set from SHCs. The overview of SHC and its applications is discussed in the next section.

3. Overview of Soil Health Cards (SHCs)

3.1 SHC in Developing Countries.

In many of the developing countries, the farmers’ or crop growers are still unaware of the soil conditions, nutrient status, what to grow, when and how? A large number of the farming population is still doing farming by using their conventional methods and have limited knowledge about soil-specific crop cultivation. A good portion of the farming community across the developing countries are doing
agriculture production based on their experience gained over the years without knowing the current condition of the soil and the means of its improvement. This unawareness leads to imbalanced fertilization that in turn causes the deficiency essential macro, secondary, and micronutrients in soils.

To enhance awareness about the soil health, fertility status, the Govt. of India (GoI) has started a national level mission in the year 2014 under soil health management (SHM) scheme for sustainable agriculture growth in India. As an extension of SHM, the Indian Agriculture and Farmer Welfare ministry has launched a scheme for providing the report card of the soils of their agriculture farms to farmers [4]. The report card of the soil is termed as Soil Health Card (SHC). Fig. 1 presents a visual depiction of an SHC which is publicly available online. The SHM and SHC based schemes are now implemented across all the states and UTs of the country through their respective state agriculture departments where SHCs are publically available in soft form and also provided to each registered farmer in hard copy form.

The SHCs are being used in many applications such as evaluation of soil fertility index, generation of soil fertility maps, and soil survey. The hard copy of the SHC in printed form is provided to each farmer for each of his/her farmland holdings and softcopy of the same SHC is available for public access. Besides the detailed information on N, P, K as macro-nutrients, S as secondary- nutrient, Zn, Fe, Cu, Mn, Bo (micronutrients) and physical parameters (PH, EC, OC), the SHC also consists of crop-specific fertilizer recommendations for the dosage to be given to soil to ensure sustainable productivity. The SHCs help the farmers in improving productivity through the judicious use of fertilizer inputs.
4. Motivation of SHC Provision

The motivation behind providing the SHCs is to enhance the knowledge of farmers about soil quality, nutrient state, soil, and crop-specific fertilizer dosage using 3R (right time, right place and right amount) concept. Since the SHCs are the outcome of the field-specific soil test, therefore the provision of SHC aims at making the farmers aware of balanced fertilization that helps them in achieving higher crop yields at optimum cost.

The SHCs are the detailed report cards of farm’s soil’s health as it gives a detailed report of health indicators and associated descriptive terms from which the health of the soil can be assessed with necessary repair action. The report cards include information on (a) soil fertility and health status, (b) crop and farm-specific fertilizer dosage, (c) recommendations on right prescription on integrated nutrient management (INM), (d) amendments of soils (saline or alkaline soil). The SHC is a very informative tool that contains a lot of information for analysis that can be useful in various agricultural applications like integrated nutrient and soil health management [7]. The information contained in cards can be collected, analyzed, and utilized effectively, can result in the best analytical results and can be used in the development of intelligent agricultural management and decision support systems. The SHC can also help in determining the changes in soil conditions over some time. The availability of SHCs helps the farmers in balanced fertilization and also reduce undue expenditure on fertilizer and other supplements.

A good quantity of SHCs having hundreds or thousands of cards in CSV or XML or pdf format can be utilized to make a good information-rich dataset for intelligent soil analysis research and development of soil specific DSS [12] or fertilizer recommendation system. This next section shall discuss the proposed method for dataset preparation in detail.

5. Proposed Method for Soil Dataset Preparation

5.1 Methodology

SHCs help farmers in assessing the impacts of their fertilizer administration choices on soil productivity and its wellbeing. The information in SHCs can be utilized to generate better decision support for farmers. For building such systems, the information in SHCs can be used to build datasets. Fig. 2 depicts the various steps of the proposed methodology for preparation of agricultural soil dataset for soil specific research and intelligent analytic applications.

![Figure 2 Steps in the Proposed Methodology](image_url)
The stepwise details of proposed methodology are:

a) Collection of SHCs in XML format
   The methodology starts with collection of hundreds or thousands of SHCs from public domain in XML data format. The SHCs in XML format acts as the main source of data for preparation of dataset.

b) Identification of Features
   Post collection, the relevant tags from SHC’s were identified as features for the resultant dataset.

c) Creation of Configuration file as per Identified features and Dataset Output format
   A configuration file is created to keep the identified features. The file can be used to add/remove any feature from dataset. The usage of configuration files gives flexibility in the creation of new datasets with different features. The same configuration file also contains the output format for the dataset.

d) Parsing of Data Source [8]
   The data source i.e. the XML files of collected SHCs are parsed using XML parser to get the relevant information from the XML files in accordance with the selected features.

e) Transformation in Configured Output Data Format
   The parsed information from the XML files is transformed into the output data format as per the details configured in configuration file.

5.2 Implementation
The main component of the proposed method is an XML parser that parses the XML files of collected SHCs into relevant information fields. Since the parser is developed using python language, hence it utilizes the open-source ElementTree library in tandem with other common data parsing libraries to parse the data from XML files. The efficient APIs of the python library helps in parsing and navigation of the XML document followed by its breakdown into ease to use a tree structure [8]. The following steps have been used iteratively to generate dataset in the commonly used data formats. Post parsing, another fast, flexible, and easy to use python library named Panda library [9] is used to build or prepare the dataset. This library is an open-source tool developed for data analysis and manipulation. For ease of use, the dataset can be generated in multiple common data formats such as .csv, xls, and, ariff, etc.

Step-1: Read the file path of SHCs extracted from internet sources
Step-2: Parse the file path
Step-3: Get all the children of the root of the tree structure
Step-4: With the tree structure representation of file, the navigation begins from the root of the tree i.e. the <data> element, containing the entire data structure.
Step-5: Iterative approach is used to iterate through each node in the tree
Step-6: Through the iterative approach, all the features with their names, their attributes, and all of the sub-elements were extracted to build the data set.

6. Experimentation, Results and Discussion
For experimentation, multiple SHCs were collected in XML file format from the public domain for a specific village in one of the states in India. One of the SHC is given in Fig.1 that was issued to a farmer by gram panchayat of that village. The corresponding XML file is depicted in Fig. 3. Similarly, multiple SHCs were taken and saved in a specific location on a personal computer. The parser program discussed in the previous section reads and parse hundreds or even more numbers of files stored in the specified location and builds dataset using other supporting python libraries as discussed in the previous section. The raw dataset thus obtained is cleaned through the removal of NaN, duplicates, missing values, removal of outliers and the final dataset is made ready to use in any data analytic application. The collection of XML files of SHCs can be either downloaded manually for a specific site or location.
The final generated soil dataset is presented in Fig. 4 where various physical and nutrient-related properties are considered as features for the dataset. For the representation purpose, the soil micro and macro nutrients along with chemical properties are selected as features. The researchers can identify and select feature as per research needs and accordingly dataset value may be obtained as dataset.

Figure 3 XML generated from the SHC via Proposed Method

Figure 4 Prepared Soil Dataset using Proposed Method
7. Conclusion and Future Scope

The soil dataset preparation method presented in the paper is quite simple, flexible and acts as a powerful dataset generator for soil quality evaluation, prediction, and other usages such as interpreting the impacts of fertilizer administration choices on well-being of soil fertility and agricultural soil conservation strategies [10]. The method can be used to build dataset for futuristic soil nutrient and productivity research on a broader scale.

The usage of the open-source, fast, and powerful libraries of python programming language makes the implementation less complex and easy to use. Since the SHCs are publically available and contains very rich site-specific nutrient details and associated crop-specific recommendation, the method shows that these SHCs in a set can be used to develop an informative dataset which can be utilized for any intelligent data analytic applications on soil research such as soil health management, soil quality assessment [11], development of intelligent soil and crop-specific DSS for fertilization recommendation using ML and AI techniques. The dataset developed by the proposed method can be utilized in any futuristic soil health-related research for the improvement of soil quality index concerning the Indian context and its interpretation.

The method initializes with manually collected XML files of SHCs. The manual collection can be replaced with an automated scraper tool that can scrap the SHCs from the online site. Scraped SHC data can also be used to generate a dataset for analysis of soil and creating prediction models to predict nutrients. This can be taken as a future scope of work for researchers

References

[1] El-Ramady HR, Alshaal TA, Amer M, Domokos-Szabolcsy É, Elhawat N, Prokisch J, Fári M. Soil quality and plant nutrition. InSustainable Agriculture Reviews 14 2014; pp. 345-447. Springer, Cham.

[2] http://soilhealth.dac.gov.in.

[3] Niranjan HK, Chouhan RS, Sharma HO, Rathi D. Awareness and performance of soil health card scheme in central India. Journal of Crop and Weed. 2018;14(1):99-103.

[4] Reddy A. Impact study of soil health card scheme. Reddy A Amarender (2017) Impact Study of Soil Health Card Scheme, National Institute of Agricultural Extension Management (MANAGE), Hyderabad-500030. 2018 Feb 1:210.

[5] Patel A, Patel P, Sipai SA. Significance of Soil Health Card on Agriculture Development. PROGRESS. 2019 Jun 8:146.

[6] Padarian J, Minasny B, McBratney AB. Machine learning and soil sciences: A review aided by machine learning tools. Soil. 2020;6(1):35-52.

[7] Young M. The technical writer's handbook: writing with style and clarity. University Science Books; 2002.

[8] Purakayastha TJ, Pathak H, Kumari S, Biswas S, Chakrabarty B, Padaria RN, Kamble K, Pandey M, Sasmal S, Singh A. Soil health card development for efficient soil management in Haryana, India. Soil and Tillage Research. 2019 Aug 1;191:294-305.

[9] Lam TC, Ding JJ, Liu JC. XML document parsing: Operational and performance characteristics. Computer. 2008 Sep 9;41(9):30-7.

[10] McKinney W. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. "O'Reilly Media, Inc."; 2012 Oct 8.
[11] Gura I, Mnkeni PN. Dataset on the use of the Soil Management Assessment Framework (SMAF) in evaluating the impact of conservation agriculture strategies on soil quality. Data in brief. 2019 Feb;22:578.

[12] Islam, S.D.U., Bhuiyan, M.A.H., Mohinuzzaman, M., Ali, M.H. and Moon, S.R., 2017. A Soil Health Card (SHC) for soil quality monitoring of agricultural lands in south-eastern coastal region of Bangladesh. Environmental Systems Research, 6(1), p.15.

[13] Motia S, Reddy S. Conceptual Framework of a Prototype Data Driven Decision Support System for Farmland Health Assessment using Wireless Sensor Network. In 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON) 2019 Mar 13 (pp. 215-222). IEEE.