Computers and Electronics in Agriculture 104 (2014) 93–104

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Machine learning assessments of soil drying for agricultural planning

Evan J. Coopersmith [a,*], Barbara S. Minsker [a], Craig E. Wenzel [b], Brian J. Gilmore [b]

[a] Department of Civil and Environmental Engineering, University of Illinois, Urbana-Champaign, Urbana, IL 61801, United States
[b] John Deere Technological Innovation Center, Champaign, IL 61820, United States

ARTICLE INFO

Article history:
Received 22 March 2013
Received in revised form 25 February 2014
Accepted 9 April 2014

Keywords:
Soil drying
Field readiness
Machine learning
Nearest neighbors
Soil moisture
Decision support

ABSTRACT

The hydrologic processes of wetting and drying play a crucial role in agricultural activities involving heavy equipment on unpaved terrain. When soil conditions moisten, equipment can become mired, causing expensive delays. While experienced users may assess soil conditions before entering off-road areas, novice users or those who must remotely assess sites before traveling may have difficulty assessing conditions reliably. One means of assessing dryness is remotely-monitored *in situ* sensors. Unfortunately, land owners hesitate to place sensors due to monetary costs, complexity, and sometimes infeasibility of physical visits to remote locations. This work addresses these limitations by modeling the wetting/drying process through machine learning algorithms fed by hydrologic data – remotely assessing soil conditions using only publicly-accessible information. Classification trees, k-nearest-neighbors, and boosted perceptrons deliver statistical soil dryness estimates at a site located in Urbana, IL. The k-nearest-neighbor and boosted perceptron algorithms both performed with 91–94% accuracy, with most misclassifications falling within calculated margins of error. These analyses demonstrate that reasonably accurate predictions of current soil conditions are possible with only precipitation and potential evaporation data. These two values are measured throughout the continental United States and are likely to be available globally from satellite sensors in the near future. Through this type of approach, agricultural management decisions can be enabled remotely, without the time and expense of on-site visitations or extensive ground-based sensory grids.

1. Introduction

Previous work has made forays into soil drying assessments over a diverse set of geographic locations, climate conditions, and functional objectives. The primary dynamic process affecting soil drying is precipitation (Entekhabi and Rodriguez-Iturbe, 1994). For this reason, models of wetting and drying have often focused upon an "antecedent precipitation index" (API), using a pre-set window of previous rainfall to estimate current levels of soil moisture (Saxton and Lenz, 1967). This particular concept of calculating an API has been applied in a variety of contexts: in conjunction with microwave sensing for soil moisture estimation (Blanchard et al., 1981), soil water recession modeling for agriculture (Choudhury and Blanchard, 1983), and for weather prediction (Wetzel and Chang, 1988). Another approach is the development of a stochastic model to estimate soil moisture distributions using daily rainfall and an initialization of soil moisture values (Farago, 1985). However, both the API and stochastic approach require an initial condition for soil moisture at the location where estimates are desired. This hampers applicability at many locations that do not have soil moisture sensors.

Other models have taken a hydrologic approach, employing precipitation and surface radiation to estimate soil moisture (Capehart and Carlson, 1994), but these models require boundary conditions, initial conditions, and parameters of a thermal and/or hydraulic nature that can be difficult to obtain broadly. Pan et al. (2003) and Pan (2012) addressed this concern by deriving a "diagnostic soil moisture equation" from a stochastic, linear partial differential equation. Soil moisture then becomes a function of a temporally-decaying sum of previous rainfall. Their approach no longer requires an initial condition, nor recalibration, but does require a soil moisture sensor at the location in question to calibrate the equation initially. Measuring soil moisture directly is plausible, but soil heterogeneity may necessitate numerous sensors to address spatial variation of soil moisture adequately (Pan and Peters-Lidard, 2008). The alternative approach of a soil water balance can be applied, but must be recalibrated frequently, since errors are cumulative (Jones, 2004).

In the agricultural arena, Gamache et al. (2009) developed a soil drying model, but its predictions require data from cone penetrometer and soil moisture sensors, two data sources that are not

* Corresponding author. Tel.: +1 610 639 2087.
E-mail address: ecooper2@gmail.com (E.J. Coopersmith).

http://dx.doi.org/10.1016/j.compag.2014.04.004
0168-1699/

currently available at most remote sites. Another inquiry along similar lines uses knowledge of soil types, which is theoretically public, but then continues to require soil moisture levels from proximal sensors (Chico-Santamarta et al., 2009).

Other approaches eschew the strategic placement of soil sensors in favor of modeling tire slip as a function of tractor properties (Sahu and Raheman, 2008) or other details such as vehicle type, speed, load distribution, number of passes, etc. (Pytka, 2009 and Lamande and Schjonning, 2008). These vehicle-specific properties are often unavailable outside of research studies. Another early work attempts to assess the suitability of site conditions, but uses very specific information that is not likely to be available to most applications, such as "stress–strain rate relations" or the results of triaxial tests (Sharifat and Kushwaha, 2000).

Lee and Wang (2009) focus solely on radar and other remote sensing data, but consider only the properties of snow coverage along with the hardness and density of mixtures of snow and ice. These traits are not appropriate for warmer weather conditions of interest to agriculture. Sliva et al. (2009) modeled drying properties of sugarcane fields in Brazil, but their model requires that the prediction occur in a well-specified location within a pre-determined time frame.

Alternatively, considerable prior research seeks techniques for improving agricultural conditions (Tullberg et al., 2007; Shoop et al., 2002 and Lebert et al., 2006) or minimizing the effects of traffic (Raper, 2005) rather than delivering a dryness assessment. For instance, one paper recommends a protocol for improving drying via the application of manure (Mosaddeghi et al., 2000).

This paper addresses these gaps by developing and testing a machine learning model of soil drying that requires only precipitation and potential evaporatranspiration estimates. Precipitation is widely available at high temporal resolution on a 1 km by 1 km grid from NEXRAD[1] throughout the continental United States. Potential evapotranspiration is available publically in Illinois from the Illinois State Water Survey, and can be estimated in other locations using three approaches (Jensen et al., 1990). The first method requires only air temperature and day length (Thornthwaite, 1948; Hamon, 1963). The second method requires air temperature and net radiation (Priestley and Taylor, 1972). The third, and most detailed approach, requires the information from the second as well as wind speed and relative humidity (Monteith, 1965). The Illinois Climate Network (ICN) data used in this analysis employs the third approach, but one of these three approaches should be applicable anywhere throughout the United States.

For the purposes of this analysis, the notion of dryness represents a user-defined assessment with qualitatively consistent designations for a particular application. For example, diverse agricultural activities (crops, livestock, etc.) may possess different notions of acceptable soil conditions, but provided the algorithm is given training data consistent to one particular context, it will adapt appropriately. This current analysis focuses upon a general test case for agricultural soil drying, where "dry" implies that a given tract of farmland is viable for a particular type of work (e.g., planting, crop treatment, or harvesting) on a given day.

Section 2 presents the case study, a brief overview of the geography and relevant features of the South Farms test site that is the focus of the data analysis presented in this work. Next, Section 3 describes the methodology used in remotely estimating dryness from public data sources. Section 4 then presents the results of the case study application and compares the relative performance of the algorithms. Finally, the paper concludes in Section 5 with an assessment of which machine learning techniques have performed most successfully, a discussion of these results in the context of previous work, and a brief discussion of potential future enhancements and other applications of this research.

2. Case study – South Farms, Urbana, IL

The methods developed in this study are tested at the University of Illinois South Farms located in Urbana IL, which is classified as a continental or microthermal climate, Dfa by the Koppen–Geiger classification system (Koppen, 1936; updated by Peel et al., 2007). The specific climate zone is characterized by a warm, humid summer and colder, drier winters. Annual rainfall levels, gathered from 1990 to 2011 at the ICN sensor located near the test site (Fig. 2.1), average approximately 1013 mm per year. The potential evapotranspiration estimate over the same time period is 1046 mm per year. The warmest month is July and the coolest is January, with average daily temperatures of 75.1 and 26.9°F (24.0 and −2.8 Celsius) respectively. As precipitation levels and potential evapotranspiration are both highest during the summer (the middle of growing season), the flat landscape yields a test site that will be characterized by multiple periods of wetting and drying during any growing season. Agricultural sites in this region are often tile-drained, which results in a shorter soil drying system memory than similar locations without the tile drains.

Fig. 2.1 presents the location of the test site, located within the Energy and Biosciences Institute (EBI) energy farm. Also pictured is the ICN sensor platform used in this analysis. The ICN sensors provide readings of potential evaporation (which incorporates solar radiation, humidity, wind, temperature, etc.) and precipitation. To the southeast are the largest plots maintained by EBI, upon which soil condition assessments were gathered.

A John Deere intern, Jordan Pitcher, provided assessments of soil conditions throughout the growing season within the green square. Mr. Pitcher has extensive agricultural experience and his assessments served as the soil dryness training and validation data for the machine learning algorithms. The sensors labeled "EBI" provide precipitation information.

3. Methodology

This section describes the framework developed for assessing dryness based on soil drying. An overview of the approach is first provided, followed by a discussion of the various input data sources, a description of the algorithms used to assess soil drying and their outputs, and concluding with the computational tools and requirements for implementation.

The first approach, the k-nearest-neighbor (KNN) algorithm, which was introduced by Fix and Hodges (1951) and deployed in many water resources and hydroinformatics applications (e.g., Kumar et al., 2006, Meliker et al., 2008, McRoberts et al., 2007; Nemes et al., 2008 and Coopersmith et al., 2011), is an intuitively satisfying approach for classification, analysis, and forecasting. The algorithm simply uses current precipitation and potential evapotranspiration measurements to locate the most similar examples from historical data (whose field conditions are known) and, in turn, leverages those similar examples to estimate the current field readiness.

The second algorithm, decision trees (also referred to as classification or regression trees), are non-parametric classification tools that recursively split datasets by values of the independent variables to minimize entropy in each subset and, thus, maximize information gain (Breiman et al., 1984). These algorithms are available in most statistical programming packages (Breiman et al., 1993) and have been deployed in a variety of environmental contexts, such as sustainable forest resource management (Aertsen

[1] http://nmq.ou.edu/beta/q2-tools.html, provided through the University of Oklahoma.