Geoderma 299 (2017) 1–12

Contents lists available at ScienceDirect

Geoderma

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

Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates

Divan Vermeulen [a], Adriaan Van Niekerk [a,b,*]

[a] Department of Geography & Environmental Studies, Stellenbosch University, Private Bag X1, Matieland, Stellenbosch 7602, South Africa
[b] School of Plant Biology, University of Western Australia, 35 Stirling Hwy, Crawley, WA, 6009 Perth, Australia

ARTICLE INFO

Article history:
Received 10 August 2016
Received in revised form 8 March 2017
Accepted 11 March 2017
Available online 30 March 2017

Keywords:
Salinity
Hydrology
Digital terrain analysis
Geomorphometry
Machine learning
Geostatistics

ABSTRACT

Conventional methods of monitoring salt accumulation in irrigation schemes require regular field visits to collect soil samples for laboratory analysis. Identifying areas prone to salt accumulation by means of geomorphometry (i.e. terrain analyses using digital elevation models (DEMs)) can potentially save time and costs. This study evaluated the extent to which DEM derivatives and machine learning (ML) algorithms (k-nearest neighbour, support vector machine, decision tree (DT) and random forest) can be used for predicting the location and extent of salt-affected areas within the Vaalharts and Breede River irrigation schemes of South Africa. In accordance with local management policies, salt-affected areas were defined as regions with soil electrical conductivity (EC) values >4 dS/m. Two DEMs, namely the one-arch second Shuttle Radar Topography Mission (SRTM) DEM and a photogrammetrically-extracted digital surface model (DSM), were used for deriving the derivatives. Wetness indices as well as hydrological and morphometric terrain analysis techniques were used to generate predictive variables. For comparative purposes, the predictive variables were also used as input to regression modelling and kriging with external drift (KED). Thresholds were applied to the regression models and KED results to obtain a binary classification. EC values based on in situ soil samples were used for model development, classifier training and accuracy assessment.

The results show that KED achieved the highest overall accuracy (OA) in Vaalharts (79.6%), whereas KED and ML (DT) showed the most promise in the Breede River (75%). The findings suggest that the use of elevation data and its derivatives as input to geostatistics and ML holds much potential for monitoring salt accumulation in irrigated areas, particularly for simulating sub-surface conditions. More work is needed to investigate the potential of using ML and DEM-derivatives, along with other geospatial datasets such as satellite imagery (that have been shown to be effective for monitoring surface conditions), for the operational modelling of salt accumulation in large irrigation schemes.

1. Introduction

Salinity is a term used to describe the amount of salt in soil or water (Mcghie and Ryan, 2005). For the purpose of this study, salinity refers to the accumulation of soluble salts in the soil due to natural processes or human activities (Al-Khaier, 2003). The way in which salts move and accumulate in soils can be affected by poor drainage (waterlogging), irrigation practices, clearance of vegetation and the reshaping of the landscape through earth works (Mcghie and Ryan, 2005). In large quantities, salts limit the normal growth of plants and the negative impacts of salt accumulation on crop production is a global concern (Metternicht and Zinck, 2003).

An estimated 18% of the soils in South African irrigation schemes is salt-affected or waterlogged (Backeberg et al., 1996). Although this percentage is relatively small compared to Argentina (34%), Egypt (33%), Iran (30%), Pakistan (26%) and the United States of America (23%) (Ghassemi et al., 1995), only 13.7% of South Africa's land area is suitable

Abbreviations: CNBL, channel network base level; CD: NGI, Chief Directorate: Dational Geo-spatial Information; CK, co-kriging; CV, coefficient of variation; CSC, cross sectional curvature; DT, decision tree; DEM, digital elevation model; DSM, digital surface model; DDG, downslope distance gradient; EC, electrical conductivity; HAND, height above nearest drainage; *k*NN, *k*-nearest neighbour; KED, kriging with external drift; ML, machine learning; MSP, mid-slope position; NH, normalized height; OK, ordinary kriging; RF, random forest; RK, regression kriging; RM, regression modelling; RSP, relative slope position; SRTM, shuttle radar topography mission; SH, slope height; SLFA, slope limited flow accumulation; SVM, support vector machine; SAGA, system for automated geostatistics analyses; SWI, SAGA wetness index; TST, terrain surface texture; TPI, topographic position index; TWI, topographic wetness index; UK, universal kriging; VDTCN, vertical distance to channel network.

* Corresponding author at: Department of Geography & Environmental Studies, Stellenbosch University, Private Bag X1, Matieland, Stellenbosch 7602, South Africa.
E-mail address: avn@sun.ac.za (A. Van Niekerk).

http://dx.doi.org/10.1016/j.geoderma.2017.03.013
0016-7061/

for irrigated agriculture (Department of Agriculture, Forestry and Fisheries, 2013). Proactive measures to reduce the effect of salt accumulation are therefore needed to prevent loss of productive agricultural land. Preventative measures include careful consideration of crop water requirements and irrigation water quality, as well as frequent monitoring of salt levels in soils (Shainberg and Shalhevet, 1984).

Conventional methods of monitoring salt-affected soils require regular field visits and laboratory analyses, which is often not viable for frequent monitoring of large areas. Although there has been an increase in the use of proximal (in situ) sensors (Viscarra Rossel et al., 2011), such instruments normally monitor soil conditions within relatively small ranges (within 2 m). This necessitates the incorporation of a large number of sensors to effectively monitor extensive areas at the required (i.e. within field) spatial resolutions. Owing to its ability to observe large areas on a regular, timely basis, remote sensing has also been used as an alternative method for monitoring salt accumulation (Abbas et al., 2013; Akramkhanov et al., 2011; Dwivedi, 1997; Dwivedi et al., 1999; Elnaggar and Noller, 2010; Sulebak et al., 2000). However, a major challenge of using remotely sensed imagery is its inability to effectively monitor subsurface processes that do not directly influence the spectral responses of the topsoil (Vermeulen and Van Niekerk, 2016).

The use of geomorphometry – terrain analysis using digital elevation data (Pike, 2000) – to model areas that are susceptible to salt accumulation has produced good results. Elnaggar and Noller (2010) found a significant correlation between soil electrical conductivity (EC), and elevation, slope and wetness indices. Similarly, Sulebak et al. (2000) identified a strong, significant correlation ($R^2 = 0.8$) between terrain data (slope, aspect and profile curvature) and soil moisture using a stepwise regression modelling (RM) approach. Sulebak et al. (2000) observed that low slope gradients were associated with high soil wetness values and Akramkhanov et al. (2011) found significant correlations (as determined by stepwise multiple regression) between soil EC and environmental factors such as distance to drainage, profile curvature, slope and groundwater table depth. Taghizadeh-mehrjardi et al. (2016) found wetness indices, the multi-resolution valley bottom flatness index and elevation to be the most important predictors of soil salinity.

Geostatistics have widely been used in salt accumulation studies (Eldeiry and Garcia, 2009, 2008; Gallichand et al., 1992; Juan et al., 2011; Li et al., 2007; Taghizadeh-Mehrjardi et al., 2014; Utset et al., 1998), particularly for interpolating salt accumulation from soil sample analysis results. Kriging, a generic term used to refer to a group of generalized least-squares regression algorithms, has been shown to produce good results, as it provides linear unbiased estimates and weights surrounding sample points to account for clustering (Gallichand et al., 1992; Hengl et al., 2007). Several variations of the kriging algorithm are available, but co-kriging (CK), universal kriging (UK), regression kriging (RK) and kriging with external drift (KED) seem to be the most popular for salt accumulation modelling (Baxter and Oliver, 2005; Bishop and McBratney, 2001; Eldeiry and Garcia, 2008; Gallichand et al., 1992; Li et al., 2007; Motaghian and Mohammadi, 2011; Taghizadeh-Mehrjardi et al., 2014).

CK, the simplest of these algorithms, is a multivariate extension of kriging that allows for the incorporation of auxiliary data to improve predictive capacity (Wackernagel, 2010). CK is suitable when only a few auxiliary variables are being considered and when these variables do not cover all sample locations (Hengl et al., 2003). UK, RK and KED are mathematically equivalent algorithms that make use of auxiliary variables to compute the kriging trend model (Pebesma, 2006). UK models the trend using coordinates only, whereas KED makes use of other auxiliary variables for estimating the trend function. RK calculates the drift and residuals separately, after which the results are summed (Hengl et al., 2007). Gallichand et al. (1992) found CK to produce better EC models compared to moving average methods, while Eldeiry and Garcia (2008) observed that RK produced a stronger model compared to those generated with RM. Performing RK, Taghizadeh-Mehrjardi et al. (2014) observed a moderate significant correlation ($R^2 = 0.49$) between soil EC and the evaluated variables, with wetness indices, geomorphological surfaces (rock outcrops), principal components, catchment aspect and valley depth being the main predictors. Li et al. (2007) showed that CK and RK produced better results than ordinary kriging (OK), emphasising the importance of incorporating ancillary data (e.g. terrain analysis derivatives) in the interpolation of EC. Comparing OK, RK and KED, Bishop and McBratney (2001) found KED to be the best predictor of soil EC, while Motaghian and Mohammadi (2011) demonstrated that KED produced more accurate results in modelling soil saturated hydraulic conductivity than RM, OK, CK and RK. Similarly, Baxter and Oliver (2005) found that KED produced superior results (compared to CK and RK) in predicting potentially available nitrogen within agricultural fields.

In contrast to geostatistical methods, machine learning (ML) algorithms use samples of known identity (categories) to classify instances of unknown identity (Campbell, 2006; Rees, 2001). Various ML algorithms, including *k*-nearest neighbour (*k*NN) (Coopersmith et al., 2014; Nemes et al., 2006, 1999), artificial neural networks (Aitkenhead et al., 2012; Behrens et al., 2005), support vector machine (SVM) (Kovacevic et al., 2010; Li et al., 2013), decision tree (DT) (Bui and Moran, 2001; Jafari et al., 2014) and random forest (RF) (Heung et al., 2014), accompanied by auxiliary variables, have been employed to predict soil properties and classes. Evans et al. (1996a) produced reasonable accuracies (>60%) for mapping saline soils with decision trees (DTs). Similar observations were made by Evans et al. (1996b). Also employing DTs for salt accumulation mapping, Elnaggar and Noller (2010) achieved very accurate results (60% and 98.8% for unaffected and salt-affected soils respectively) and attributed it to the algorithm's ability to incorporate a large number of disparate predictors in the model building process. DTs are, however, prone to overfitting (i.e. producing models that perform well on the training data, but poorly on general untrained data), while more powerful machine learning algorithms such as SVM and RF have been shown to be more robust (Rodriquez-Galiano et al., 2012a; Rodriquez-Galiano et al., 2012b; Myburgh and Van Niekerk, 2014).

Although much work has been done on combining ML algorithms and remotely sensed imagery for mapping salt-affected areas (Abbas et al., 2013; Abbas and Khan, 2007; Abood et al., 2011; Dwivedi and Sreenivas, 1998; Elnaggar and Noller, 2010; Muller and Van Niekerk, 2016; Vermeulen and Van Niekerk, 2016), such data can only observe surface conditions. The use of digital elevation models (DEMs) (and its derivatives) as input to ML algorithms for delineating salt-affected areas is of particular interest, as it would better represent subsurface conditions. However, we are not aware of any published studies in which ML algorithms were compared to other established methods (e.g. geostatistics) when only terrain variables were used as input. In addition, very little information is available on the impact of DEM properties on salt accumulation modelling.

This study aims to evaluate the use of several ML algorithms (*k*NN, SVM, DTs and RFs) for identifying areas in irrigated fields that are salt-affected. The main purpose is to determine the effectiveness of these methods for producing simple binary maps of salt-affected and unaffected areas so that they can be used as a scoping mechanism to prioritize more detailed (in situ) investigations and to discard unaffected areas from further consideration. The ML results are compared to binary classifications applied to models generated by two established methods, namely RM and KED. The Vaalharts and Breede River irrigation schemes in South Africa (Figs. 1 and 2) were chosen as study sites. The landscapes of the two areas are very different, with Vaalharts mostly consisting of flat terrain, while Breede River is located in a mountainous region. This allowed for a better comparison and evaluation of the techniques.