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.

© 2017 Elsevier B.V. All rights reserved.

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/© 2017 Elsevier B.V. All rights reserved.

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.

2. Materials and methods

2.1. Study areas

The Vaalharts irrigation scheme (Fig. 1) is located in the Northern Cape Province of South Africa in the Harts River valley, close to the small towns Hartswater and Jan Kempsdorp. The irrigation scheme's altitude ranges from 1064 m to 1154 m above mean sea level and it covers an area of 36,950 ha. Vaalharts has a semi-arid climate, with a mean annual rainfall of 400 mm (Schulze et al., 2006), cold winters and long warm summers (Barnard et al., 2012). Most rain occurs during the summer months (November to April) when the mean monthly rainfall is 47.3 mm and the mean minimum and maximum temperatures are 4.4 °C and 38.8 °C respectively. During the winter months, the

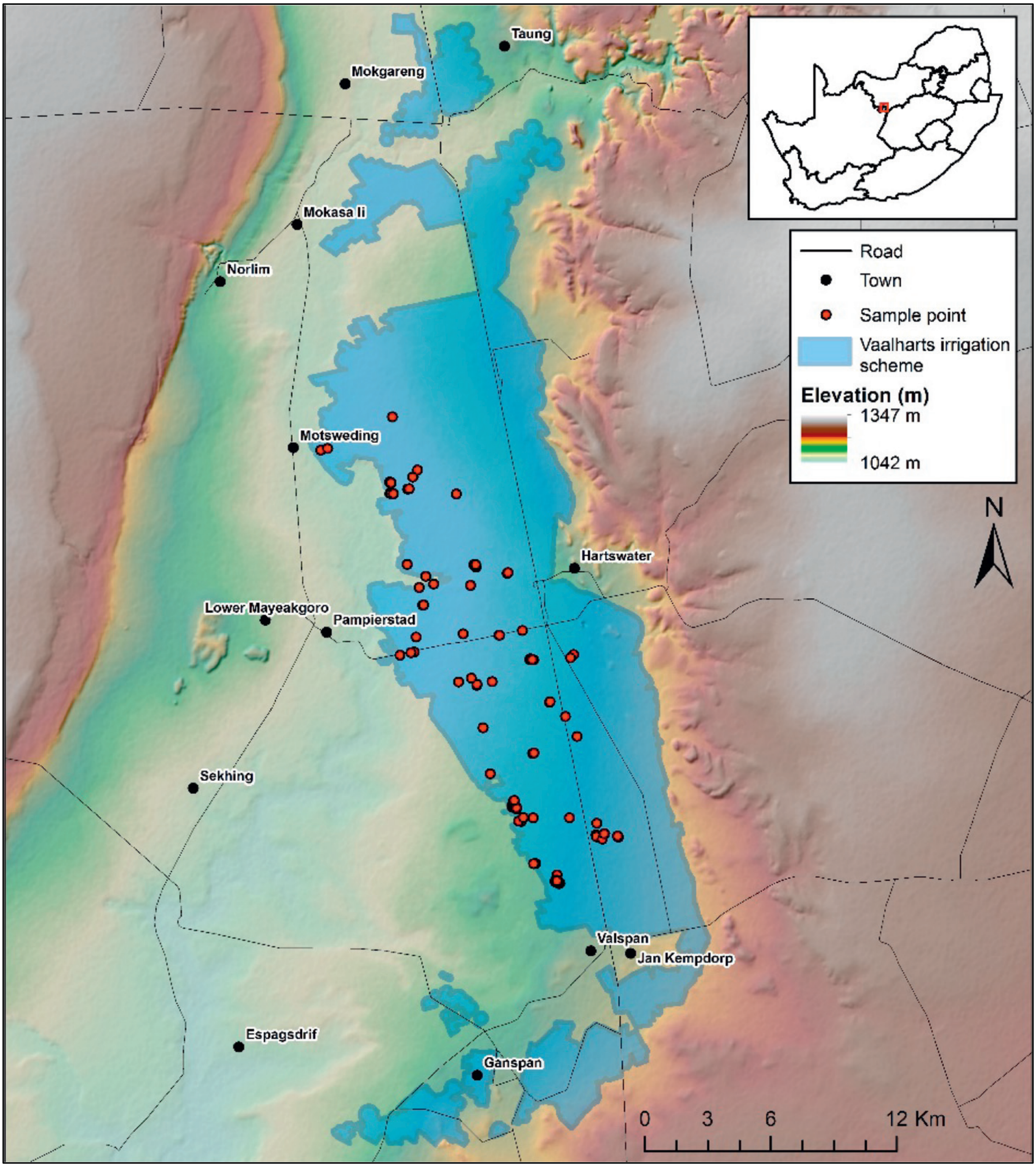

Fig. 1. Location of the Vaalharts study area. Also shown is the extent of the Vaalharts irrigation scheme and the distribution of the soil samples obtained during the field surveys.

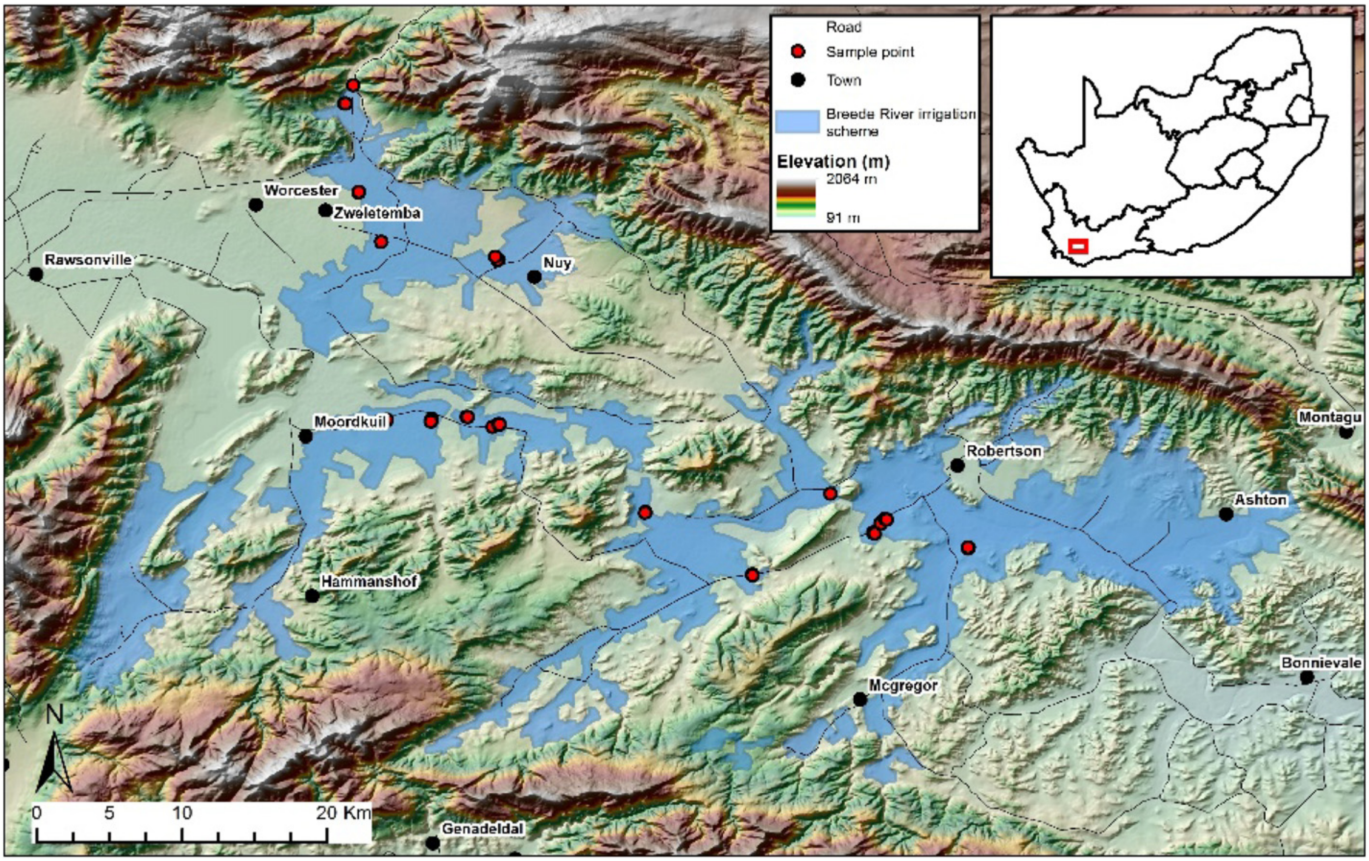

Fig. 2. Location of the Breede River study area. Also shown is the extent of the Breede River irrigation scheme and the distribution of the soil samples obtained during the field surveys.

mean monthly rainfall drops to 4.1 mm and the mean minimum and maximum temperatures are −4.8 °C and 31.8 °C respectively (Schulze and Maharaj, 2006).

The Vaalharts irrigation scheme mostly consists of sandy soils, but is prone to waterlogging and salt accumulation due to insufficient natural drainage (Maisela, 2007). The dominant salts in the area are Ca and Mg (HCO_3). Soils in the scheme typically consist of 8% clay, 2% silt, 68% fine sand and 22% medium and course sand (Streutker, 1977). Maize, wheat, barley, lucerne and groundnuts are the main crops (Kruger et al., 2009) and are mostly planted on a rotational basis (Maisela, 2007).

Sloping towards the south, the Harts River valley is bordered by two plateaus, one to the east and one to the west. Due to the low gradient of the Harts River and an absence of incisions by the river (Barnard et al., 2012; Gombar and Erasmus, 1976), the topography of valley is unvarying (Fig. 1). Vaalharts consists of the Archaean Ventersdorp Supergroup and older basement rocks composed of Archaean Kraaipan Group sediments and volcanic rock. Most of the study area is overlaid by Rietgat and Allanridge Formations, both of the Ventersdorp Supergroup. The eastern region of the Harts River valley comprises the Allanridge Formation, which includes basalt and andesite. The Rietgat Formation includes sandstone, tuff, limestone and andesitic lavas (Liebenberg, 1977). To the west of the valley, the Schimidtdrif and Camvellrand subgroups can be found, which consist of dolomite, limestone, quartzite, shale and chert (Schutte, 1994).

The Breede river irrigation scheme (Fig. 2) is located in Breede River valley in the Western Cape province of South Africa and covers an area of 57,415 ha. Its altitude ranges from 91 m to 2064 m above mean sea level. Nearby towns include Robertson and Worcester. The area has a Mediterranean climate and a mean annual rainfall of 324 mm. Rainfall mostly occurs during the winter months (between May and October), when the mean monthly rainfall is 31 mm (Schulze et al., 2006). During these months, the mean minimum and maximum temperatures are 1.3 °C and 30.7 °C respectively (Schulze and Maharaj, 2006). In summer the mean minimum and maximum temperatures increase to 4.3 °C and 37.6 °C respectively (Schulze and Maharaj, 2006), while the mean monthly rainfall drops to 10.8 mm (Schulze et al., 2006). Vines, orchards and lucerne are the main crops planted in the area. Due to the inflow of saline irrigation from various irrigation districts, the quality of the irrigation water in the Breede River progressively deteriorates in a downstream, south-eastern direction. This creates problems for downstream users extracting water from the river for agricultural irrigation (Ghassemi et al., 1995).

The central region of the Breede River valley has a gentle hilly topography and lies between the Langeberg Mountains in the north and the Riviersonderend Mountains in the south. The Langeberg Mountains reach heights of about 2064 m, whereas heights of the Riviersonderend Mountain range between 300 m and 800 m. These mountain ranges consist of sandstone and resistant quartzites to the north, and sandstone and shales to south (Ghassemi et al., 1995; Kirchner, 1995). The mountain ranges composed mostly of quarzites comprise Hutton, Clovelly, Constantia, La Motte and Champagne soils, whereas pediments and valley floors consist of the Clovelly, Constantia, La Motte, Fernwood and Champagne soils. Irrigated soils along the Breede River may also be composed of Dundee, Oaklead and Fernwood soils (Lambrechts, 1979). The dominant salts in Breede River region are Na and Cl (NaCl).

2.2. Data collection and preparation

Stereoscopic aerial photographs, the Shuttle Radar Topography Mission (SRTM) DEM and soil samples were collected for this study. Details about each dataset are provided in the following subsections.

2.2.1. Digital elevation models (DEMs)

The 30 m resolution SRTM DEM was acquired for both study areas (Rabus et al., 2003). The SRTM DEM covers the entire globe and is freely available. The DEM was produced from C-band radar (Hensley et al., 2000) and has a horizontal accuracy of 20 m and an absolute vertical accuracy of about 16 m (Smith and Sandwell, 2003). According to Guth (2006) and Hayakawa et al. (2008), the SRTM DEM underestimates slope in high-relief areas and overestimates average slope in flat regions. More information on the accuracy and resolution of the SRTM DEM can be found in Hayakawa et al. (2008) and Huggel et al. (2008).

Stereoscopic aerial images acquired from the Chief Directorate: National Geo-spatial Information (CD: NGI) of South Africa was used to develop a 2 m spatial resolution digital surface model (DSM) of each study area. The DSM generation comprised two steps, namely 1) epipolar pair generation; and 2) image matching. During epipolar pair generation, the y-parallax between left and right stereo images was removed to reduce the processing time of identifying matching pixels during the image matching step (Deilami and Hashim, 2011; Zomer et al., 2002). Image matching was then used to automatically identify matching features (points, lines, curves and regions) on the overlapping stereo images. Elevation values were then extracted to produce high quality DSMs (Zhang and Fraser, 2008). PCI Geomatica 2013 software was used for generating the DSMs in this study.

Given that groundwater flow tends to follow general topographic patterns and will therefore depend less on small-scale variations (Sørensen and Seibert, 2007), the resolution of the generated DSMs was resampled to 20 m using the Aggregate tool in ArcGIS 10. A cell factor of four and the mean aggregation type was employed. Thompson et al. (2001) found that lowering a DEM's horizontal resolution reduces local variance (noise) and results in a product that is smoother and more suitable for terrain analyses.

2.2.2. Input variable generation

The System for Automated Geoscientific Analyses (SAGA) software package was used to generate the following DEM derivatives (Akramkhanov et al., 2011; Elnaggar and Noller, 2010; Taghizadeh-mehrjardi et al., 2016; Taghizadeh-Mehrjardi et al., 2014): 1) elevation; 2) slope aspect; 3) slope gradient; 4) convergence index; 5) cross sectional curvature (CSC); 6) longitudinal curvature; 7) relative slope position (RSP); 8) mid-slope position (MSP); 9) normalized height (NH); 10) slope height (SH); 11) standardized height; 12) downslope distance gradient (DDG); 13) real surface area; 14) terrain ruggedness index; 15) terrain surface texture (TST); 16) topographic position index (TPI); 17) channel network base level (CNBL); 18) closed depressions; 19) LS-factor (slope length and steepness factor); 20) valley depth; 21) vertical distance to channel network (VDTCN); 22) catchment area; 23) slope limited flow accumulation (SLFA); 24) topographic wetness index (TWI); 25) SAGA wetness index (SWI) (Böhner et al., 2006, 2002); and 26) height above nearest drainage (HAND) (Rennó et al., 2008). All 26 variables were derived from the SRTM DEM and DSMs respectively, resulting in two feature sets per study area.

A principal component analysis (PCA), which condenses variables to produce a set of uncorrelated variables ordered in terms of variance (Behrens et al., 2010; Eldeiry and Garcia, 2009), was applied to each feature set. Nearly all (99.9%) of the variance of the input variables were condensed to a single component for both the DSM and SRTM DEM derivatives in Vaalharts. In Breede River, the first principle component contained 99.9% of the variance associated with the DSM derivatives, while containing 85.6% of SRTM DEM derivatives. The resulting first principle component (PC1) was included in each feature set. Table 1 summarizes the derivatives (27) considered in this study.

2.2.3. Soil sample collection

A total of 175 and 63 soil samples were collected for the Vaalharts and Breede River study areas respectively. The sample size was based on the suggestion of Beleites et al. (2013) that a minimum of 25 samples per class is needed to assess the performance of a classifier. A comparatively larger number of samples were collected in the Vaalharts area owing to the large range of annual crops that is being produced on a rotation basis throughout the year. Samples were clustered in and around areas where clear indications of salt accumulation occurred (e.g. visible salt crusts on the soil surface). Unaffected soil samples were also collected in the immediate neighbourhood. Topsoil samples were collected at a depth of 0–20 cm by means of a soil auger. The sample positions were measured using a differential global positioning system (GPS) with an accuracy of 10 cm. The EC was determined in a laboratory using the saturated paste technique described by the Soil Society of South Africa (1991).

Table 1
DEM derivatives included in each feature set.

Type	Variables[a]	# of variables
Elevation	Mean height above sea level, NH, standardized height	3 × 2
Hydrology	CNBL, closed depressions, LS-factor, valley depth, VDTCN, catchment area, SLFA, HAND	8 × 2
Morphometry	Aspect, slope, convergence index, CSC, longitudinal curvature, RSP, MSP, SH, DDG, real surface area, terrain ruggedness index, TST, TPI	13 × 2
Wetness indices	TWI, SWI	2 × 2
Image transformations	PC1	1 × 2

[a] NH, normalized height; CNBL, channel network base level; LS-factor, slope length and steepness factor; VDTCN, vertical distance to channel network; SLFA, slope limited flow accumulation; HAND, height above nearest drainage; CSC, cross-sectional curvature; RSP, relative slope position; MSP, mid-slope position; SH, slope height; DDG, downslope distance gradient; TST, terrain surface texture; TPI, topographic position index; TWI, topographic wetness index; SWI, SAGA wetness index; PC1, first principal component.

2.3. Analyses

A range of 27 terrain variables were generated from two different DEMs and used as input to the ML classifiers. For comparison purposes, the same data was used as input to RM and KED. The results of the experiments were interpreted in the context of finding an operational solution for monitoring salt accumulation in large irrigated areas. Given the large number (1062 in total) of experiments that were carried out, only those that provided noteworthy results are reported and discussed in this manuscript.

2.3.1. Regression modelling

RM, as implemented in IBM SPSS v20.0 software, was used to statistically analyse the relationship between the soil EC and the DEM derivatives. Linear, logarithmic, inverse, quadratic, cubic, power and exponential regression models were evaluated. Stepwise multiple regression and partial least squares (PLS) were also carried out on the 27 input variables (Cho et al., 2007; Hansen and Schjoerring, 2003).

2.3.2. Geostatistics

By considering a continuous attribute (z) at any unsampled location (u) using z-data ($\{z(u_\alpha), \alpha = 1, \ldots, n\}$), the basic linear regression estimator ($Z^*(u)$) for all kriging algorithms can be defined as (Goovaerts, 1999):

$$Z^*(u)-m(u)=\sum_{\alpha=1}^{n(u)}\beta_\alpha(u)[Z(u_\alpha)-m(u_\alpha)]$$

where $\beta_\alpha(u)$ is the weight assigned to datum $z(u_\alpha)$ interpreted as a realization of the random variable $Z(u_{\alpha})$ and is located within a given neighbourhood $W(u)$ centred on u. To minimize error variance, the

$n(u)$ weights are chosen under the constraint of unbiasedness of the estimator. Variants in kriging methods are dependent on the trend $m(u)$ variable of the algorithm. Intended as a generalized case of kriging, KED models the trend as a function of the available auxiliary data (Hengl et al., 2007). The $m(u)$ trend for KED is calculated as follows (Goovaerts, 1999):

$$m(u') = \sum_{k=0}^{K} a_k(u') f_k(u')$$

where $\alpha_k(u') \approx a_k$ constant but unknown $\forall\ u' \in W(u)$.

A more detailed description of kriging can be found in Goovaerts (1999).

KED was performed on each individual variable as well as on the full set (27). Subsets of derivatives were also evaluated to investigate whether a reduction in feature space dimensionality improves the results. The subsets consisted of the two, three and four individual derivatives with the highest overall accuracies (OAs). In addition to the non-logarithmic KED models, logarithmic transformations were also applied to each of the models to reduce the effect of skewed distributions and very large values, as was the case with the EC values obtained in this study (Gundogdu and Guney, 2007). Several variograms (e.g. exponential, quadratic), which defines the variations between neighbouring values as a function of the geographic distance between the evaluated points within a study area (Eldeiry and Garcia, 2009), were employed.

Two types of variogram approaches to estimate a grid from a set of points, namely global-fit and local-fit, were used. The latter only takes into account the significant sample points identified for a selected area within the study area, whereas global-fit calculates a single function for the entire study area (Gundogdu and Guney, 2007). The SAGA software package was used to perform the global-fit KED.

A threshold value of 4 dS/m was applied to the regression and geostatistical models to produce a binary classification of salt-affected and unaffected areas. Modelled values greater than this threshold were considered salt-affected. A sensitivity analysis was carried out to investigate whether the use of different thresholds (from 2 to 6 dS/m) would have any marked impact on the classification results.

2.3.3. Machine learning

The *k*NN, SVM, DT and RF ML classifications were performed using the OpenCV implementations of the algorithms (Bradski, 2000). *k*NN is a simple non-parametric, distance-based classifier that labels each unknown instance based on its *k* neighbouring known instances (Cover and Hart, 1967; Gibson and Power, 2000). *k*NN has the disadvantage of assigning equal weight to all variables, even though certain variables may have higher priority. This can result in incorrect class assignments and diffuse clusters (Cunningham and Delany, 2007). To avoid this, only odd k-values (namely 1, 3 and 5) were used in this study, as suggested by Campbell (2006).

SVM determines the optimal separating hyperplane between classes by focussing on the training samples (support vectors) close to the edge of the class descriptors and consequently minimizing misclassifications (Lizarazo, 2008; Novack et al., 2011; Tzotsos and Argialas, 2006). As recommended by Hsu et al. (2010), the kernel type for the SVM classifier was set to the radial basis function. The remaining parameters were left as default. See Vapnik (2000) and Huang et al. (2002) for a more detailed explanation of SVM.

A DT identifies relationships between a response variable known as the dependent variable, and multiple, continuous variables known as the independent variables. DTs hierarchically split a dataset into increasingly homogeneous subsets known as nodes (Gómez et al., 2012; Punia et al., 2011). By recursively splitting the feature datasets, a leaf node is reached, with the class associated with the node assigned to the observation (Pal and Mather, 2003). According to Pal and Mather (2003) and Novack et al. (2011), each node is limited to a split in feature space orthogonal to the axis of the selected feature. Each branch of the DT consists of divisions (or rules) of the most probable class. Applying these rules will assign the most likely class to an unknown instance (Lawrence and Wright, 2001).

RF is an enhancement of DTs (Immitzer et al., 2012) and generates each DT by using a random vector sampled independently from the input vector. A vote is cast by each of the generated DTs (Bosch et al., 2007; Breiman, 2001; Pal, 2005). Each classifier contributes a single vote to the assignment of the most popular class of the input variable (Breiman, 2001; Rodriquez-Galiano et al., 2012a). RF makes use of bagging (Breiman, 2001; Rodriquez-Galiano et al., 2012a), a method which generates a training set for feature selection. This allows RF classifiers to have a low (even lower than DT classifiers) sensitivity to training set size (Rodriquez-Galiano et al., 2012a). Two parameters are required to be set, namely the number of trees and the number of active (predictive) variables. The number of active variables for RF was set to one, three, five and ten, whereas the number of trees was set to 100. Rodriquez-Galiano et al. (2012a) showed that stability in accuracy is achieved at 100 trees and that a small number of split variables are optimal for reducing generalization errors and correlations between trees. A more detailed discussion of the RF classifier can be found in Breiman (2001, 1996), Pal (2005) and Rodriquez-Galiano et al. (2012b).

A total of 125 and 43 (70% of total) soil samples were used for training the classifiers in the Vaalharts and Breede River study areas respectively, while the rest of the samples were used for accuracy assessment.

2.4. Accuracy assessment

Maps were created from the rule-based and supervised classifications to identify salt-affected areas within the study areas. An independent set of 50 and 20 soil samples were used as reference samples in the Vaalharts and Breede River study areas respectively. Confusion matrices were used to calculate the overall accuracy (OA), producer accuracy (PA), user accuracy (UA), kappa coefficient and the area under receiver operating characteristic (AUROC) curve (Congalton and Green, 2009; Evangelista, 2006).

3. Results

Fig. 3 shows that a good balance between salt-affected and unaffected samples was achieved. More dramatic differences in salinity levels were noted during the field surveys for Vaalharts compared to Breede River. Samples consisting of EC values of <4 dS/m were classified as unaffected, whereas samples with EC measurements equal to or >4 dS/m

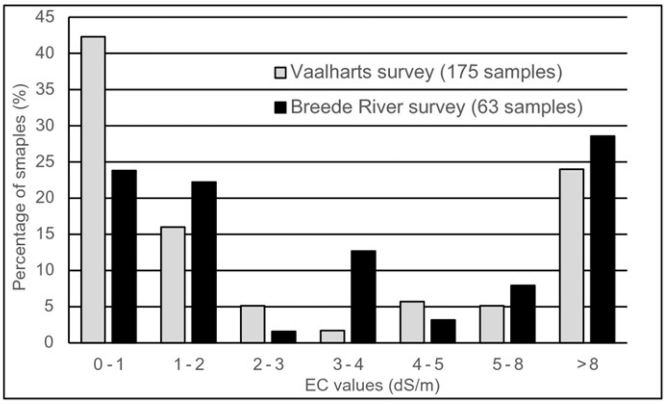

Fig. 3. EC values of soil samples collected in the study areas.

Table 2
Statistical profiles of the training and reference samples for each study area.

Measurement (dS/m)[a]	Vaalharts			Breede River		
	All samples	Training	Reference	All samples	Training	Reference
Mean	9.7	9.5	10.2	12.2	9.5	18.2
Median	1.5	1.3	3	3.2	3.2	2.9
Minimum	0.1	0.1	0.1	0.4	0.4	0.6
Maximum	95	95	84	81.6	81.6	80.8
CV	2.1	2.1	1.9	1.7	1.8	1.4
Kurtosis	7.4	7.4	8.4	4.4	9.3	1.1
Skewness	2.8	2.8	2.9	2.3	3	1.5

[a] CV, coefficient of variance.

were considered to be salt-affected (Soil Society of South Africa, 1991; Nell and Van Niekerk, 2014).

Table 2 shows the mean, median, minimum, maximum, coefficient of variation (CV), kurtosis and skewness of the training and reference samples for each of the study areas. From Table 2 it is clear that there is sufficient relation between the training and reference samples.

Table 3 summarizes the accuracies of the RM, KED and ML classifications. For the sake of brevity, only the three strongest models for each method are included in the table. Also provided are the source DEM and the feature set on which the method was performed. For the regression models the goodness-of-fit (R^2) values are noted.

3.1. Regression modelling

SH derived from the SRTM DEM produced the strongest model ($R^2 = 0.71$, $p < 0.001$) in Vaalharts, with the relationship being best described by a cubic model. A scatterplot of this model is provided in Fig. 4. From the scatterplot it is clear that RM tends to underestimate the EC for highly saline samples, with a large proportion of highly saline samples being modelled as having near-zero dS/m values. Weak regression models were produced in the Breede River, with normalized height generating the best model ($R^2 = 0.15$, $p < 0.001$). Weaker models were produced from the stepwise multiple and the PLS regression for both Vaalharts ($R^2 < 0.6$) and Breede River ($R^2 < 0.1$). The kappa values of the best performing regression models suggest a "fair agreement" with the reference data (Landis and Koch, 1977).

The model achieved an OA of 68% (kappa = 0.36) when classified. The marginally weaker ($R^2 = 0.68$, $p < 0.001$) quadratic model produced a slightly better classification (OA = 72%; kappa = 0.44). In spite of the poor goodness-to-fit of this model, it was still reasonably successful (OA = 65%; kappa = 0.3) in separating salt-affected from unaffected soils when classified. As explained in Section 2.3.2, different thresholds (from 2 to 6 dS/m) for classifying salt-affected and unaffected areas were considered to assess model sensitivity. The results (not shown here) were consistent with those when 4 dS/m was used as threshold. In some cases, overall accuracies did improve slightly, but at the expense of an imbalance between the user's and producer's accuracies.

Table 3
Three top performing regression models, KED models and ML classifiers for the study areas.

Method	Study area[a]	DEM	Feature[b]	Model	R^{2} [c]	Class	PA (%)	UA (%)	OA (%)	Kappa	AUROC[d]
Regression modelling	VH	SRTM	SH	Quadratic	0.68^{*}	Salt-affected	72	72	**72**	**0.44**	**0.72*****
						Unaffected	72	72			
	VH	SRTM	SH	Cubic	0.71^{*}	Salt-affected	76	65.5	68	0.36	0.68^{***}
						Unaffected	60	71.4			
	VH	SRTM	DDG	Cubic	0.65^{*}	Salt-affected	40	50	50	0	0.5^{***}
						Unaffected	60	50			
	BR	DSM	NH	Power	0.15^{**}	Salt-affected	60	66.7	**65**	**0.3**	**0.65*****
						Unaffected	70	63.6			
	BR	DSM	VD	Exponential	0.12^{**}	Salt-affected	20	50	50	0	0.5^{***}
						Unaffected	80	50			
	BR	SRTM	RSP	Quadratic	0.11^{**}	Salt-affected	100	50	50	0	0.5^{***}
						Unaffected	70	63.6			
Kriging	VH	DSM	CSC,ASP,STDH	KED		Salt-affected	91.7	73.3	**79.6**	**0.59**	**0.8*****
						Unaffected	68	89.5			
	VH	SRTM	TST	KED		Salt-affected	100	68.6	77.6	0.55	0.78^{***}
						Unaffected	56	100			
	VH	DSM	CSC, ASP	KED		Salt-affected	91.7	68.8	75.5	0.51	0.76^{***}
						Unaffected	60	88.2			
	BR	SRTM	NH	KED		Salt-affected	90	69.2	**75**	**0.5**	**0.75*****
						Unaffected	60	85.7			
	BR	SRTM	DDG	KED		Salt-affected	80	72.7	75	0.5	0.75^{***}
						Unaffected	70	77.8			
	BR	DSM	TRI, VD	KED		Salt-affected	80	72.7	75	0.5	0.75^{***}
						Unaffected	70	77.8			
Machine learning	VH	SRTM & DSM	All	DT		Salt-affected	56	73.7	**68**	**0.36**	**0.68*****
						Unaffected	80	64.5			
	VH	DSM	All	RF (a = 5)		Salt-affected	40	76.9	64	0.28	0.64^{***}
						Unaffected	88	59.5			
	VH	SRTM	All	kNN (k = 1)		Salt-affected	56	66.7	64	0.28	0.64^{***}
						Unaffected	72	62.1			
	BR	DSM	All	DT		Salt-affected	90	69.2	**75**	**0.5**	**0.75*****
						Unaffected	60	85.7			
	BR	SRTM	All	DT		Salt-affected	70	70	70	0.4	0.7^{***}
						Unaffected	70	70			
	BR	DSM	All	RF (a = 5)		Salt-affected	50	62.5	60	0.2	0.6^{***}
						Unaffected	70	58.3			

Note: The bold values represents the best performing model for each method.

[a] VH, Vaalharts; BR, Breede River.

[b] SH, slope height; DDG, downslope distance gradient; NH, normalized height; VD, valley depth; RSP, relative slope position; CSC, cross-sectional curvature; ASP, aspect; STDH, standardized height; TST, terrain surface texture; TRI, terrain ruggedness index.

[c] *, regression results significant at a 0.001 level; **, regression results significant at a 0.01 level.

[d] ***, AUROC results significant at a 0.05 level.

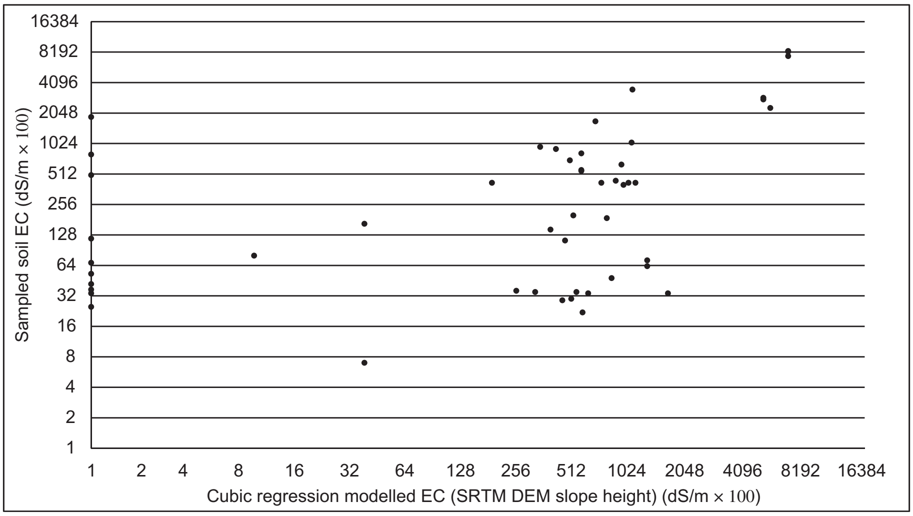

Fig. 4. Scatterplot of Vaalharts SRTM DEM cubic regression model derived from slope height.

3.2. Geostatistics

The exponential algorithm was chosen as the appropriate variogram for both the non-logarithmic and logarithmic KED models for Vaalharts and for the logarithmic Breede River KED model, while the quadratic algorithm was found to best represent the non-logarithmic model for Breede River. The non-logarithmic KED model combining the DSM derived cross-sectional curvature, aspect and SH variables achieved the highest OA (79.6%) for Vaalharts (kappa = 0.59). When the scatterplot of this model is interpreted (Fig. 5a), it is clear that the relationship between the modelled and measured EC is erratic. The KED model also tends to underestimate a large number of highly saline samples as non-saline. The SRTM derived normalized height model (non-logarithmic) showed the most promise for Breede River, producing the highest accuracy (OA = 75%; kappa = 0.5). According to the kappa values of these models, there is a "moderate agreement" with the reference data (Landis and Koch, 1977). Interestingly, a logarithmic transformation of the input data produced lower OA compared to the untransformed models for both Vaalharts (OA < 65%) and Breede River (OA < 70%). As with the RM model for Vaalharts, a number of highly saline samples were incorrectly classified as being non-saline by the model.

As with the RM, the KED models were largely insensitive to variations in classification thresholds (from 2 to 8 dS/m).

3.3. Machine learning

The DT classifier achieved the highest accuracy in both Vaalharts (OA = 68%; kappa = 0.36) and Breede River (OA = 75%; kappa = 0.5). The former model was based on both the SRTM DEM and the DSM variables, while in the latter only DSM derivatives were required to achieve the highest accuracies. RF (based on DTs) also attained higher accuracies than the other classifiers, achieving an OA of 64% (kappa = 0.28) in Vaalharts and an OA of 60% (kappa = 0.5) in Breede River. Both RF classifications consisted of the DSM feature set.

3.4. Classified maps

Fig. 6 and Fig. 7 shows the thematic maps of the top four performing classifications for Vaalharts and Breede River respectively. The KED classifications in Vaalharts (Fig. 6a to Fig. 6c) appear to be very similar, whereas a substantial difference in the distribution of modelled salt-affected areas is observed when the quadratic regression model (based on SH) classification (Fig. 6d) is considered.

In Breede River the maps of the KED models (Fig. 7a to Fig. 7c) are less similar, but as with the Vaalharts, the KED models predict large, continuous salt-affected regions. The DT classification (based on the DSM derivatives) resulted in smaller patches of salt accumulation.

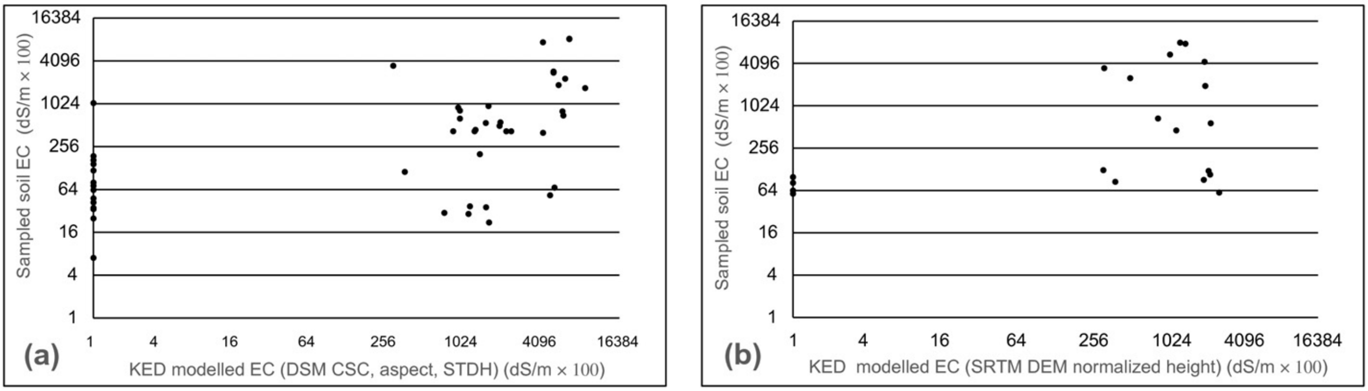

Fig. 5. Scatterplots of (a) Vaalharts KED model produced from cross-sectional curvature, aspect and standardized height and (b) Breede River SRTM DEM KED model derived from normalized height.

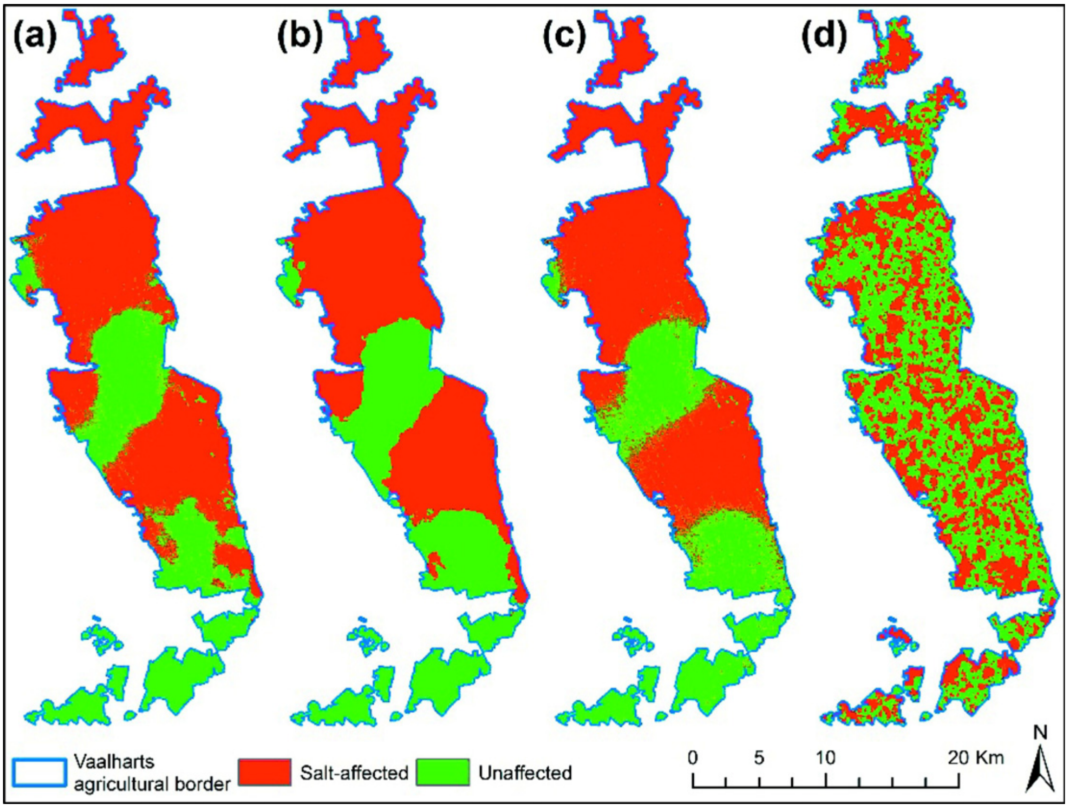

Fig. 6. Maps produced from the (a) DSM KED model derived from cross-sectional curvature, aspect and standardized height, (b) SRTM DEM KED model derived from terrain surface texture, (c) DSM KED model derived from cross-sectional curvature and aspect, and (d) quadratic regression model produced from slope height for Vaalharts.

4. Discussion

In spite of its superior resolution, there was little difference between the DSM-based results and those generated from the SRTM DEM. This can be attributed to the influence of land cover features (e.g. vegetation), particularly in the Breede River where the crops are mainly perennial and woody (e.g. fruit trees and vineyards) (Ghassemi et al., 1995). Improvements in accuracies were observed in some cases (especially in Vaalharts) where both the SRTM and DSM derivatives were used as input to the ML classifiers, which suggests that the data source did have an influence on the results. Better results may have been obtained with using a digital terrain model (DTM) generated from light detection and ranging (LiDAR) data, since it has the ability to penetrate foliage (Hesse, 2010), but to obtain such data for large regions is still prohibitively expensive, especially in developing countries such as South Africa.

KED produced the most consistent results (OA and kappa standard deviation of 1.75% and 0.03 respectively) and attained the highest accuracies (mean OA and kappa of 76.28% and 0.53 respectively) throughout, which suggest that this technique is most suitable for modelling salt accumulation when only DEM derivatives are used as input. This result is in agreement with Douaoui and El Ghadiri (2015), Douaoui et al. (2006), Eldeiry and Garcia (2009), Gallichand et al. (1992), Juan et al. (2011) and Taghizadeh-Mehrjardi et al. (2014). Although these studies focussed on either OK, CK or RK, the high OA of the KED classification in the present study demonstrates its potential for modelling salt-affected soils. Our results also support the findings of Bishop and McBratney (2001), who applied KED to elevation and terrain data to model soil EC.

A major advantage of KED is its ability to include more than one terrain derivative as input. However, based on our experiments, a decrease in accuracy was observed when more than three derivatives were included in the KED model. Some form of input variable selection is consequently required prior to performing KED. The iterative variable selection approach used in this study will likely be too laborious for operational implementations.

Classifying the regression models into salt-affected and unaffected areas produced relatively poor results (mean OA of 59.2%). This is attributed to the large spatial variation of salt accumulation and the inability of regression modelling to consider autocorrelation effects (Overmars et al., 2003). Spatial autocorrelation, which occurs when information from samples located near each other are not independent (Dormann et al., 2007), can have a positive or negative impact on accuracy assessment results due to the influence of errors at particular locations on neighbouring locations (Congalton, 1991).

The ML classifiers showed improvement in accuracies (mean OA = 66.8%; kappa 0.34) over the regression models, but generally attained lower accuracies than the KED models (mean OA = 76.28%, kappa = 0.53). However, in Breede River, the DT classifier (based on DSM variables) was able to match the accuracies of KED (OA = 75%; kappa = 0.5), but was unable to compete with KED in Vaalharts. All of the other ML classifiers (*k*NN, RF and SVM) performed relatively poorly (OA < 70%; kappa ≤ 0.4) in both study areas when compared to the KED models. In spite of this relatively poor performance, ML should not be disregarded, as a major advantage of ML algorithms is their ability to incorporate various types of input data, including remotely sensed imagery. Several studies have shown that ML algorithms are very effective for mapping salt affected areas using satellite imagery (Abbas et al., 2013; Abbas and Khan, 2007; Abood et al., 2011; Dwivedi and Sreenivas, 1998; Muller and Van Niekerk, 2016; Vermeulen and Van Niekerk, 2016), but many authors have noted that such data only consider surface conditions (Dwivedi, 1997; Dwivedi et al., 1999; Metternicht and Zinck, 2003). In this study we specifically focussed on using terrain data only because it is likely to better represent subsurface conditions. The fact that the DT classifier was able to match

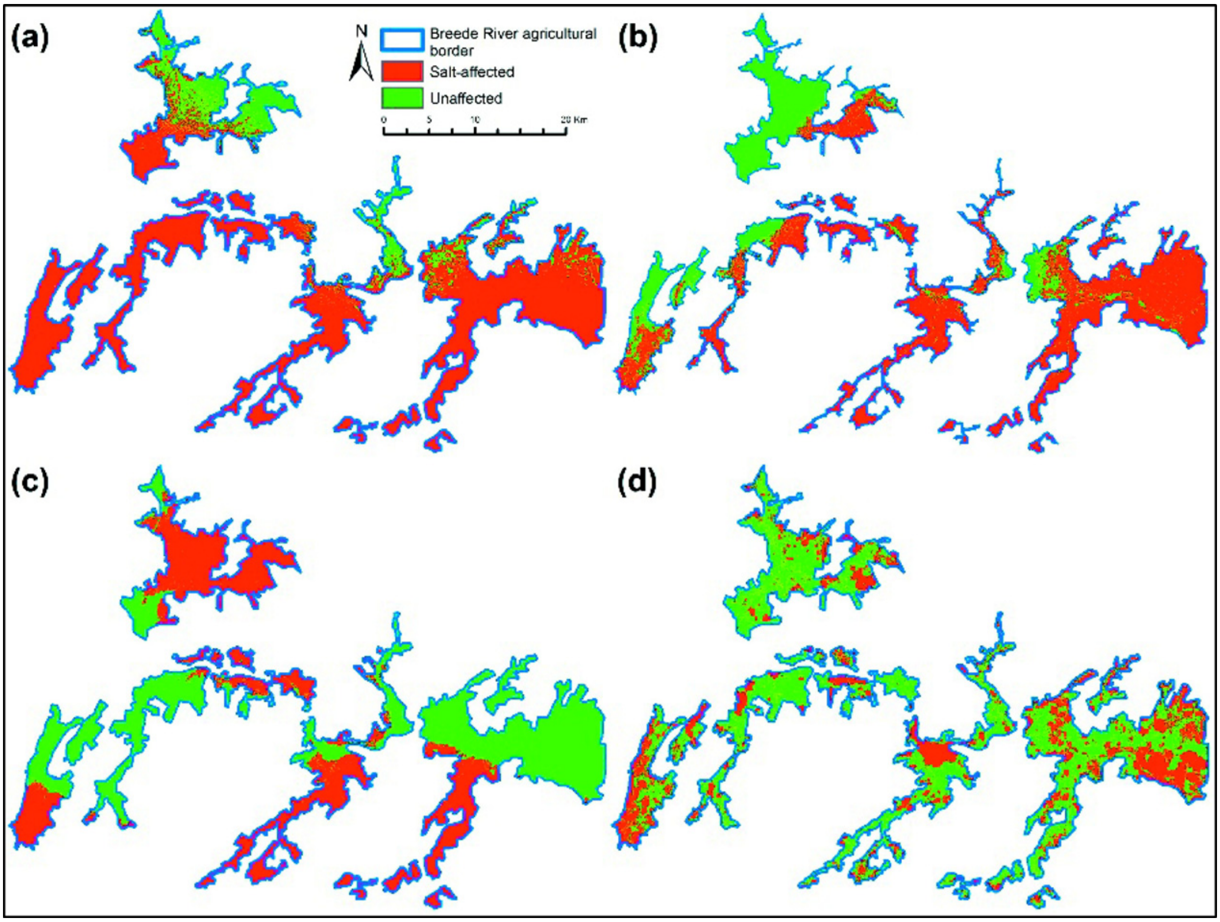

Fig. 7. Maps produced from the (a) SRTM DEM KED model derived from normalized height, (b) SRTM DEM KED model derived from downslope distance gradient, (c) DSM KED model derived from terrain ruggedness index and valley depth, and (d) DSM based DT classifier for Breede River.

the KED results in Breede River is encouraging, because it supports the findings of Elnaggar and Noller (2010) that ML can be used to investigate both surface (using remotely sensed imagery) and sub-surface (using terrain data) conditions. Another advantage of DT is that the resulting tree can be interpreted to better understand the relationships between the input variables and salt accumulation. A model based on DT can also potentially be transferred to other areas with similar conditions without the need for collecting new samples. Clearly more work is needed to investigate the value of using DT (and other ML algorithms) when terrain data is combined with remotely sensed imagery – and other geospatial datasets such as soil type maps (where available) – for the operational monitoring of salt accumulation.

A number of highly saline samples were incorrectly classified by both KED and RM as having near-zero salinity levels (see Figs. 4 and 5). During the field surveys it was observed that salt accumulation often occurred in small patches and that salinity levels varied dramatically over very short distances (<10 m). In many cases, these variations occurred on homogenous terrain and were likely caused by farming practices (e.g. over-irrigation). The modelling errors could also have been a factor of the insufficient detail of the DEMs employed, as small variations within fields were often inadequately represented. The use of more detailed DEMs (e.g. those obtained from high-density LiDAR) will likely improve results.

In this study a binary classification scheme (i.e. affected and unaffected) was adopted because the purpose was to identify salt-affected and unaffected areas so that they can be used as a scoping mechanism to prioritize more detailed (in situ) investigations. However, it would be of great value to investigate whether the techniques considered in this study would be able to differentiate more salinity classes, e.g. non-saline (<2 dS/m), slightly-saline (2–4 dS/m), moderately saline (4–8 dS/m) and strongly saline (>8 dS/m).

5. Conclusion

This study evaluated the use of RM, KED and ML techniques for identifying areas in irrigated fields where salts are likely to accumulate. The methods were evaluated in two study areas, namely the Vaalharts and Breede River irrigation schemes of South Africa. The SRTM DEM and a DSM derived from high-resolution stereoscopic aerial photographs were used as the primary data sources. A total of 27 derivatives were generated from the DEMs and used as input to the models evaluated. The results showed that KED outperformed the RM and ML classifiers in most cases, but that ML (specifically the DT classifier) was able to match KED in the Breede River. The source of elevation data did not have a marked influence on the model outputs, although the higher resolution DSM did perform better when combined with ML in the Breede River study area.

From the results of this study, it can be concluded that the use of elevation data and its derivatives, along with geostatistics and ML algorithms, hold much potential for identifying salt-affected areas in irrigated fields. More research is needed to investigate the value of using ML algorithms for classifying a combination of DEM derivatives, satellite images, proximal sensors, other geospatial datasets and a salinity classification scheme making use of multiple cut-off values. This is especially important in the context of finding operational solutions for identifying areas prone to salt accumulation, as the routine collection of large sets of training (and reference) data is not viable for large irrigation schemes, and the inability of terrain derivatives to adequately explain variations in ECe.

Acknowledgements

This work forms part of a larger project titled "Methodology to monitor the status of water-logging and salt-affected soils on selected irrigation schemes in South Africa", which was initiated and funded by the Water Research Commission (WRC) of South Africa (contract number K5/1880//4). More information about this project is available in Report TT 648/15 [ISBN 978-1-4312-0739-8], available at www.wrc.org.za. The authors also acknowledge and thank the project leader, Dr. Piet Nell of the Agricultural Research Council, for providing the soil data used in this study and for his invaluable guidance and insight, particularly during the field surveys.

References

Abbas, A., Khan, S., 2007. Using remote sensing techniques for appraisal of irrigated soil salinity. In: Oxley, L.K.D. (Ed.), Proceedings of the MODSIM 2007 International Congress on Modelling and Simulation, pp. 2632–2638.

Abbas, A., Khan, S., Hussain, N., Hanjra, M.A., Akbar, S., 2013. Characterizing soil salinity in irrigated agriculture using a remote sensing approach. Phys. Chem. Earth 56–57, 43–52.

Abood, S., Maclean, A., Falkowski, M., 2011. Soil Salinity Detection in the Mesopotamian Agricultural Plain Utilizing WorldView-2 Imagery. Michigan Technological University.

Aitkenhead, M.J., Coull, M.C., Towers, W., Hudson, G., Black, H., 2012. Predicting soil chemical composition and other soil parameters from field observations using a neural network. Comput. Electron. Agric. 82:108–116. http://dx.doi.org/10.1016/j.compag.2011.12.013.

Akramkhanov, A., Martius, C., Park, S., Hendrickx, J., 2011. Environmental factors of spatial distribution of soil salinity on flat irrigated terrain. Geoderma 163:55–62. http://dx.doi.org/10.1016/j.geoderma.2011.04.001.

Al-Khaier, F., 2003. Soil Salinity Detection Using Satellite Remote Sensing. (Thesis). Int. Inst. Geo-information Sci. Earth Obs. Michigan Technological University.

Backeberg, G., Bembridge, T., Bennie, A., Groenwald, J., Hammes, P., Pullen, R., Thompson, H., 1996. Policy Proposal for Irrigation in South Africa. South Africa, Pretoria.

Barnard, J., Bennie, A., Du Preez, C., Sparrow, J., Van Rensburg, L., 2012. Managing Salinity Associated with Irrigation at Orange-Riet and Vaalharts Irrigation Schemes. Bloemfontein, South Africa.

Baxter, S.J., Oliver, M.A., 2005. The spatial prediction of soil mineral N and potentially available N using elevation. Geoderma 128:325–339. http://dx.doi.org/10.1016/j.geoderma.2005.04.013.

Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E., Goldschmitt, M., 2005. Digital soil mapping using artificial neural networks. J. Plant Nutr. Soil Sci. 168:21–33. http://dx.doi.org/10.1002/jpln.200421414.

Behrens, T., Zhu, A.-X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155:175–185. http://dx.doi.org/10.1016/j.geoderma.2009.07.010.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., Popp, J., 2013. Sample size planning for classification models. Anal. Chim. Acta 760:25–33. http://dx.doi.org/10.1016/j.aca.2012.11.007.

Bishop, T., McBratney, A., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103:149–160. http://dx.doi.org/10.1016/S0016-7061(01)00074-X.

Böhner, J., Köthe, R., Conrad, O., Gross, J., Ringeler, A., Selige, T., 2002. Soil regionalisation by means of terrain analysis and process parameterisation. In: Micheli, E., Nachtergaele, F., Montanarella, L. (Eds.), European Soil Bureau. The European Soil Bureau, Joint Research Centre, EUR 20398 EN, Ispra, pp. 213–222.

Böhner, J., McCloy, K., Strobl, J., 2006. SAGA: Analysis and Modelling Applications. 115th ed. (Göttinger).

Bosch, A., Zisserman, A., Muoz, X., 2007. Image Classification using Random Forests and Ferns, in: 2007 IEEE 11th International Conference on Computer Vision. Miami. : pp. 1–8 http://dx.doi.org/10.1109/ICCV.2007.4409066.

Bradski, G., 2000. The OpenCV library. Dr. Dobb's J. Softw. Tools Prof. Program. 25, 122–125.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 12–20.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24:123–140. http://dx.doi.org/10.1007/BF00058655.

Bui, E., Moran, C., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103, 79–94.

Campbell, J., 2006. Introduction to Remote Sensing. Taylor & Francis, London.

Cho, M.A., Skidmore, A., Corsi, F., van Wieren, S.E., Sobhan, I., 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. Int. J. Appl. Earth Obs. Geoinf. 9:414–424. http://dx.doi.org/10.1016/j.jag.2007.02.001.

Congalton, R., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35–46 (doi: export date 6 May 2013).

Congalton, R.G., Green, K., 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Taylor & Francis, London, England.

Coopersmith, E., Minsker, B., Wenzel, C., Gilmore, B., 2014. Machine learning assessments of soil drying for agricultural planning. Comput. Electron. Agric. 104:93–104. http://dx.doi.org/10.1016/j.compag.2014.04.004.

Cover, T., Hart, P., 1967. Nearest neighbour classification. IEEE Trans. Inf. Theory 13, 21–27.

Cunningham, P., Delany, S.J., 2007. K -nearest neighbour classifiers, technical report UCD-CSI-2007-4. Dublin. http://dx.doi.org/10.1016/S0031-3203(00)00099-6.

Deilami, K., Hashim, M., 2011. Very high resolution optical satellites for DEM generation: a review. Eur. J. Sci. Res. 49, 542–554.

Department of Agriculture, Forestry and Fisheries, 2013. Abstract of Agricultural Statistics [WWW Document]. (accessed 5.23.15). http://www.nda.agric.za/docs/statsinfo/Abstact2013.pdf.

Dormann, C.F., Mcpherson, J.M., Arau, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Ohlemu, R., Peres-neto, P.R., Schurr, F.M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography (Cop.). 30:609–628. http://dx.doi.org/10.1111/j.2007.0906-7590.05171.x.

Douaoui, A., El Ghadiri, I., 2015. Combination of remote sensing and kriging to improve soil salinity in the Hmadna plain (Algeria). Soil-Water J. 1–5.

Douaoui, A., Nicolas, H., Walter, C., 2006. Detecting salinity hazards within a semiarid context by means of combining soil and remote-sensing data. Geoderma 134, 217–230.

Dwivedi, R., Sreenivas, K., 1998. Delineation of salt-affected soils and waterlogged areas in the indo-Gangetic plains using IRS-1C LISS-III data. Int. J. Remote Sens. 19, 2739–2751.

Dwivedi, R.S., 1997. Mapping waterlogged areas in part of the Indo-Gangetic plains using remote sensing. Geocarto Int. 12, 65–70.

Dwivedi, R.S., Sreenivas, K., Ramana, K.V., 1999. Inventory of salt-affected soils and waterlogged areas: a remote sensing approach. Int. J. Remote Sens. 20:1589–1599. http://dx.doi.org/10.1080/014311699212623.

Eldeiry, A., Garcia, L., 2009. Comparison of regression Kriging and Cokriging techniques to estimate soil salinity using Landsat images. Hydrol. Days 27–38.

Eldeiry, A., Garcia, L., 2008. Detecting soil salinity in alfalfa fields using spatial modeling and remote sensing. Soil Sci. Soc. Am. 72:201–211. http://dx.doi.org/10.2136/sssaj2007.0013.

Elnaggar, A.A., Noller, J.S., 2010. Application of remote-sensing data and decision-tree analysis to mapping salt-affected soils over large areas. Remote Sens. 2, 151–165.

Evangelista, P., 2006. The Unbalanced Classification Problem: Detecting Breaches in Security (Rensselaer Polytechnic Institute).

Evans, F., Caccetta, P., Ferdowsian, R., 1996a. Integrating Remotely Sensed Data with Other Spatial Data Sets to Predict Areas at Risk from Salinity. Perth, Australia.

Evans, F.H., Ferdowsian, R., Campbell, N.A., 1996b. Predicting Salinity in the Wadjekanup and Byenup Hill Catchments (Australia).

Gallichand, J., Buckland, G., Marcotte, D., Hendry, M., 1992. Spatial interpolation of soil salinity and sodicity for a saline soil in Southern Alberta. Can. J. Soil Sci. 72, 503–516.

Ghassemi, F., Jakeman, A., Nix, H., 1995. Salinization of Land and Water Resources: Human Causes, Extent, Management and Case Studies. University of New South Wales Press, Sydney.

Gibson, P., Power, C., 2000. Introductory Remote Sensing: Digital Image Processsing and Applications. Routledge, New York.

Gombar, O., Erasmus, C., 1976. Vaalharts ontwaterings projek. South Africa, Pretoria.

Gómez, C., Wulder, M.A., Montes, F., Delgado, J.A., 2012. Modeling forest structural parameters in the mediterranean pines of central Spain using QuickBird-2 imagery and classification and regression tree analysis (CART). Remote Sens. 4:135–159. http://dx.doi.org/10.3390/rs4010135.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89:1–45. http://dx.doi.org/10.1016/S0016-7061(98)00078-0.

Gundogdu, K., Guney, I., 2007. Spatial analyses of groundwater levels using universal kriging. J. Earth Syst. Sci. 116, 49–55.

Guth, P., 2006. Geomorphometry from SRTM. Photogramm. Eng. Remote. Sens. 72, 269–277.

Hansen, P.M., Schjoerring, J.K., 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. Remote Sens. Environ. 86:542–553. http://dx.doi.org/10.1016/S0034-4257(03)00131-7.

Hayakawa, Y., Oguchi, T., Lin, Z., 2008. Comparison of new and existing global digital elevation models: ASTER G-DEM and SRTM-3. Geophys. Res. Lett. 35:1–5. http://dx.doi.org/10.1029/2008GL035036.

Hengl, T., Heuvelink, G., Rossiter, D., 2007. About regression-kriging: from equations to case studies. Comput. Geosci. 33:1301–1315. http://dx.doi.org/10.1016/j.cageo.2007.05.001.

Hengl, T., Heuvelink, G., Stein, A., 2003. Comparison of Kriging with External Drift and Regression-Kriging [WWW Document]. ITC, Tech. note. URL http://www.itc.nl/library/Academic_output/ (accessed 7.10.16).

Hensley, S., Rosen, P., Gurrola, E., 2000. The SRTM topographic mapping processor. Geosci. Remote Sens. Symp. 2000. Proceedings. IGARSS 2000. IEEE 2000 Int. 3. : pp. 1168–1170 http://dx.doi.org/10.1109/IGARSS.2000.858056.

Hesse, R., 2010. LiDAR-derived local relief models: a new tool for archaeological prospection. Archaeol. Prospect. 17, 67–72.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random Forest approach. Geoderma 214–215:141–154. http://dx.doi.org/10.1016/j.geoderma.2013.09.016.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2010. A Practical Guide to Support Vector Classification [WWW Document]. http://dx.doi.org/10.1177/02632760022050997.

Huang, C., Davis, L.S., Townshend, J.R.G., 2002. An assessment of support vector machines for land cover classification. Int. J. Remote Sens. 23, 725–749.

Huggel, C., Schneider, D., Miranda, P., Delgado Granados, H., Kääb, A., 2008. Evaluation of ASTER and SRTM DEM data for lahar modeling: a case study on lahars from Popocatépetl Volcano. Mexico. J. Volcanol. Geotherm. Res. 170:99–110. http://dx.doi.org/10.1016/j.jvolgeores.2007.09.005.

Immitzer, M., Atzberger, C., Koukal, T., 2012. Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data. Remote Sens. 4: 2661–2693. http://dx.doi.org/10.3390/rs4092661.

Jafari, A., Khademi, H., Finke, P.A., Van De Wauw, J., Ayoubi, S., 2014. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. Geoderma 232–234:148–163. http://dx.doi.org/10.1016/j.geoderma.2014.04.029.

Juan, P., Mateu, J., Jordan, M., Mataix-Solera, J., Meléndez-Pastor, I., Navarro-Pedreño, J., 2011. Geostatistical methods to identify and map spatial variations of soil salinity. J. Geochemical Explor. 108:62–72. http://dx.doi.org/10.1016/j.gexplo.2010.10.003.

Kirchner, J., 1995. Investigation into the Contribution of Ground Water to the Salt Load of the Breede River, Using Natural Isotopes and Chemical Tracers. South Africa, Pretoria.

Kovacevic, M., Bajat, B., Gajic, B., 2010. Soil type classification and estimation of soil properties using support vector machines. Geoderma 154:340–347. http://dx.doi.org/10.1016/j.geoderma.2009.11.005.

Kruger, M., Van Rensburg, J.B.J., Van den Berg, J., 2009. Perspective on the development of stem borer resistance to Bt maize and refuge compliance at the Vaalharts irrigation scheme in South Africa. Crop. Prot. 28:684–689. http://dx.doi.org/10.1016/j.cropro.2009.04.001.

Lambrechts, J., 1979. Fynbos ecology: A preliminary synthesis. In: Day, J., Siegfried, W., Louw, G., Jarman, M. (Eds.), Geology. Geomorphology and Soils. South Africa, Pretoria, pp. 16–26.

Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lawrence, R.L., Wright, A., 2001. Rule-based classification systems using classification and regression tree (CART) analysis. Photogramm. Eng. Remote Sens. 67, 1137–1142.

Li, M., Im, J., Beier, C., 2013. Machine learning approaches for forest classification and change analysis using multi- temporal Landsat TM images over Huntington wildlife forest. GIScience Remote Sens. 50:361–384. http://dx.doi.org/10.1080/15481603.2013.819161.

Li, Y., Shi, Z., Wu, C., Fang, Li, Yi, H., Li, F., 2007. Improved prediction and reduction of sampling density for soil salinity by different geostatistical methods. Agric. Sci. China 6: 832–841. http://dx.doi.org/10.1016/S1671-2927(07)60119-9.

Liebenberg, L., 1977. Die geologie van die gebied 2724D (Andalusia). Bloemfontein, South Africa.

Lizarazo, I., 2008. SVM-Based segmentation and classification of remotely sensed data. Int. J. Remote Sens. 29:7277–7283. http://dx.doi.org/10.1080/01431160802326081.

Maisela, J., 2007. Realizing Agricultural Potential in Land Reform: The Case of Vaalharts Irrigation Scheme in the Northern Cape Province. University of the Western Cape.

Mcghie, S., Ryan, M., 2005. Salinity Indicator Plants. New South Wales, Australia.

Metternicht, G.I., Zinck, J.A., 2003. Remote sensing of soil salinity: potentials and constraints. Remote Sens. Environ. 85, 1.

Motaghian, H., Mohammadi, J., 2011. Spatial estimation of saturated hydraulic conductivity from terrain attributes using regression, kriging, and artificial neural networks. Pedosphere 21:170–177. http://dx.doi.org/10.1016/S1002-0160(11)60115-X.

Muller, S., Van Niekerk, A., 2016. An evaluation of supervised classifiers for indirectly detecting salt-affected areas at irrigation scheme level. Int. J. Appl. Earth Obs. Geoinf. 49: 138–150. http://dx.doi.org/10.1016/j.jag.2016.02.005.

Myburgh, G., Van Niekerk, A., 2014. Impact of training set size on object-based land cover classification: a comparison of three classifiers. Int. J. Appl. Geospatial Res. 5, 49–67.

Nell, J., Van Niekerk, A., 2014. Appropriate methods for monitoring salt accumulation and water logging on south African irrigation schemes, in: Third International Salinity Forum. Riverside, California.

Nemes, A., Lilly, A., Oude Voshaar, J., Otto, H., 1999. Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. Geoderma 90, 187–202.

Nemes, A., Rawls, W., Pachepsky, Y., 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70:327–336. http://dx.doi.org/10.2136/sssaj2005.0128.

Novack, T., Esch, T., Kux, H., Stilla, U., 2011. Machine learning comparison between WorldView-2 and QuickBird-2-simulated imagery regarding object-based urban land cover classification. Remote Sens. 3:2263–2282. http://dx.doi.org/10.3390/rs3102263.

Overmars, K., de Koning, G., Veldkamp, A., 2003. Spatial autocorrelation in multi-scale land use models. Ecol. Model. 164:257–270. http://dx.doi.org/10.1016/S0304-3800(03)00070-X.

Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26:217–222. http://dx.doi.org/10.1080/01431160412331269698.

Pal, M., Mather, P., 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sens. Environ. 86, 554–565.

Pebesma, E., 2006. The role of external variables and GIS databases in geostatistical analysis. Trans. GIS 10:615–632. http://dx.doi.org/10.1111/j.1467-9671.2006.01015.x.

Pike, R.J., 2000. Geomorphometry -diversity in quantitative surface analysis. Prog. Phys. Geogr. 24, 1–20.

Punia, M., Joshi, P., Porwal, M., 2011. Decision tree classification of land use cover for Delhi, India using IRS-P6 AWiFS data. Expert Syst. Appl. 38, 5577–5583.

Rabus, B., Eineder, M., Roth, A., Bamler, R., 2003. The shuttle radar topography mission: a new class of digital elevation models acquired by spaceborne radar. ISPRS J. Photogramm. Remote Sens. 57, 241–262.

Rees, W., 2001. Physical Principles of Remote Sensing. Cambridge University Press, New York.

Rennó, C.D., Nobre, A.D., Cuartas, L.A., Soares, J.V., Hodnett, M.G., Tomasella, J., Waterloo, M.J., 2008. HAND, a new terrain descriptor using SRTM-DEM: mapping terra-firme rainforest environments in Amazonia. Remote Sens. Environ. 112:3469–3481. http://dx.doi.org/10.1016/j.rse.2008.03.018.

Rodriquez-Galiano, V.F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P.M., Jeganathan, C., 2012a. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. Remote Sens. Environ. 121:93–107. http://dx.doi.org/10.1016/j.rse.2011.12.003.

Rodriquez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012b. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67:93–104. http://dx.doi.org/10.1016/j.isprsjprs.2011.11.002.

Schulze, R., Lynch, S., Maharaj, M., 2006. Monthly Rainfall and Its Inter-Annual Variability. South Africa, Pretoria.

Schulze, R., Maharaj, M., 2006. Daily Mean Temperatures. South Africa, Pretoria.

Schutte, I., 1994. Die geologie van die gebied Christiana: Toeligting tot blad 2724 Christiana van die Geologiese opname van Suid-Afrika. South Africa, Pretoria.

Shainberg, I., Shalhevet, J., 1984. Soil Salinity under Irrigation: Processes and Management. Springer Verlag, Berlin.

Smith, B., Sandwell, D., 2003. Accuracy and resolution of shuttle radar topography mission data. Geophys. Res. Lett. 30:1467–1470. http://dx.doi.org/10.1029/2002GL016643.

Soil Society of South Africa, 1991. Methods of Soil Analysis. SSSSA: Non-Affiliated Soil Analysis Working Committee. Pretoria, South Africa.

Sørensen, R., Seibert, J., 2007. Effects of DEM resolution on the calculation of topographical indices: TWI and its components. J. Hydrol. 347:79–89. http://dx.doi.org/10.1016/j.jhydrol.2007.09.001.

Streutker, A., 1977. The dependence of permanent crop production on efficient irrigation and drainage at the Vaalharts government water scheme. Water SA 3, 90–102.

Sulebak, J., Tallaksen, L., Erichsen, B., 2000. Estimation of areal soil moisture by use of terrain data. Geogr. Ann. 82:89–105. http://dx.doi.org/10.1111/j.0435-3676.2000.00009.x.

Taghizadeh-mehrjardi, R., Ayoubi, S., Namazi, Z., Zolfaghari, A.A., Sadrabadi, F.R., 2016. Prediction of soil surface salinity in arid region of central Iran using auxiliary variables and genetic programming. Arid L. Res. Manag. 30, 49–64.

Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., Malone, B., 2014. Digital mapping of soil salinity in Ardakan region, central Iran. Geoderma 213:15–28. http://dx.doi.org/10.1016/j.geoderma.2013.07.020.

Thompson, J.A., Bell, J.C., Butler, C.A., 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. Geoderma 100, 67–89.

Tzotsos, A., Argialas, D., 2006. Support vector machine classification for object-based image analysis. In: Blaschke, T., Lang, S., Hay, G. (Eds.), Object-Based Image Analysis: Lecture Notes in Geoinformation and Cartography. Springer, Berlin, pp. 663–677.

Utset, A., Ruiz, M., Herrera, J., Ponce de Leon, D., 1998. A geostatistical method for soil salinity sample site spacing. Geoderma 86, 143–151.

Vapnik, V., 2000. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Vermeulen, D., Van Niekerk, A., 2016. Evaluation of a WorldView-2 image for soil salinity monitoring in a moderately affected irrigated area. J. Appl. Remote Sens. 10. http://dx.doi.org/10.1117/1.JRS.10.026025.

Viscarra Rossel, R., Adamchuk, V., Sudduth, K., McKenzie, N., Lobsey, C., 2011. Proximal soil sensing: an effective approach for soil measurements in space and time. Advances in Agronomy:pp. 237–282 http://dx.doi.org/10.1016/B978-0-12-386473-4.00010-5.

Wackernagel, H., 2010. Multivariate Geostatistics: An Introduction with Applications. third ed. Springer Verlag, Berlin, Germany.

Zhang, C., Fraser, C., 2008. Generation of digital surface model from high resolution satellite imagery. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 37, 785–790.

Zomer, R., Ustin, S., Ives, J., 2002. Using satellite remote sensing for DEM extraction in complex mountainous terrain: landscape analysis of the Makalu Barun National Park of eastern Nepal. Int. J. Remote Sens. 23:125–143. http://dx.doi.org/10.1080/01431160010006449.