

Development of a Spatial Hydrologic Soil Map Using Spectral Reflectance Band Recognition and a Multiple-Output Artificial Neural Network Model

Khamis Naba Sayl^{1,3}, Haitham Abdulmohsin Afan¹, Nur Shazwani Muhammad¹, Ahmed ElShafie²

¹Department of Civil and Structural Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

²Department of Civil Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia

³Department of Dams and Water Resources, Engineering College, University of Anbar, Ramadi, Iraq

Correspondence to: Nur Shazwani Muhammad (shazwani.muhammad@ukm.edu.my)

The authors express sincere appreciation and thanked the Editorial Board and Reviewer #2 for their constructive comments and suggestions. These comments will definitely improve the quality of the manuscript. The authors have addressed all the concerns raised by Reviewer#2 in an itemized fashion below. The authors will include all suggestions in the revised version of the manuscript.

Reviewer #2:

General Comments

1) Clarity of methods section, as written, the methods section is not repeatable. Much effort is needed to more specific in describing steps taken and cite all algorithms used. Couple examples: 1- soil sampling technique (e.g. auger, open pit) and sampling depths were never specified, 2- radial basis neural network has no citations, 3- the unsupervised classification method is not specified or cited, and the software used not specified. There are quite a few major methodological details where I got lost and could not figure out what was done (please see the commented pdf for full details).

Reply:

The authors agree with the reviewer that more description on the methodology should be included in the manuscript. Therefore, the following paragraph that gives more details on soil sampling will be added to the revised manuscript:

“There are twenty five (25) sampling locations throughout the study area. The selection of these sampling locations was made based on its accessibility, vegetation cover, the spectral signatures urban areas, water, slope, roads, soil roughness, location and topography. On-site, a GPS instrument was used to locate the sampling points. The soil samples were collected using auger and the depths were between 20 to 40 cm of the topsoil. Subsequently, the soil samples were brought to the laboratory for particle size distribution analysis. This analysis was done to determine the soil texture, which is an important parameter for hydrological behavior of soil. Particles above 50 μm were separated using wet sieving techniques, while small particles were analyzed using hydrometer method (Ryan et al. 1996; Skopp 2000). Soil particles are dispersed in water and the reduction of fluid density due to setting is determined through hydrometer (Skopp 2000). The particle size distributions of all collected soil samples were classified according to the USGS method texture classification.

The following paragraphs will be added in order to provide more details on the neural network and unsupervised classification process performed in this study:

“Image classification method was used in order to extract the most relevant information needed to achieve the objectives of this study. This technique includes the process of sorting pixels into a finite number of classes or categories of data based on their data file values (Jain 1989) If a pixel satisfies a certain set of criteria, then it is assigned to the class that corresponds to that set of criteria. A pixel may be characterized by its spectral signature, which is determined by the relative reflectance in the different wavelength bands. Multispectral classification is an information-extraction process that analyzes these spectral signatures and then assigns pixels to categories based on similar signatures. One form of classification is unsupervised classification (Acharya et al. 2005). Unsupervised classification was carried out using the ERDAS Imagine software and it involves the identification of clusters or groupings in a feature space. A cluster is a set of points in the feature space where their local density is large (relative maximum) compared to the density of feature points in the surrounding region. Unsupervised classification permits an unbiased assessment of the total of the raw data. It can be used on the first hand to identify the main classes and then check the information in the field (Richards and Jia 1999). This method is preferred if area is very large and field data is lacking. It illustrates that a priori knowledge of the human

observer is needed to assign the pixels to classes. One of the most known unsupervised classifier method is the Iterative Self Organizing Data Analysis (ISODATA). ISODATA is an iterative process where it repeatedly performs an entire classification (outputting a thematic raster layer) and recalculates the statistics. Self-Organizing refers to the way in which it locates clusters with minimum user input (Acharya et al. 2005). Each iteration calculates means and reclassifies pixels with respect to the new means. Iterative class splitting, merging, and deleting are done based on the input threshold parameters. All pixels are classified to the nearest class unless standard deviation or distance threshold is specified. One of the primary advantages of unsupervised classification is many of the classes are created automatically (Jain 1989)”

The following information on Radial Basis Neural Network (RBNN) will be added to the revised manuscript:

“Radial Basis Neural Network (RBNN) is an artificial method that based on the interpolation of a multivariate function (Lowe and Broomhead 1988) . RBNN consists of three layers, i.e. input layer for feeding feature vector to the network, hidden layer where the calculation of outcome of basis function is processed, and finally the output layer for linear combining the basic functions. The following figure shows the structure of RBNN.

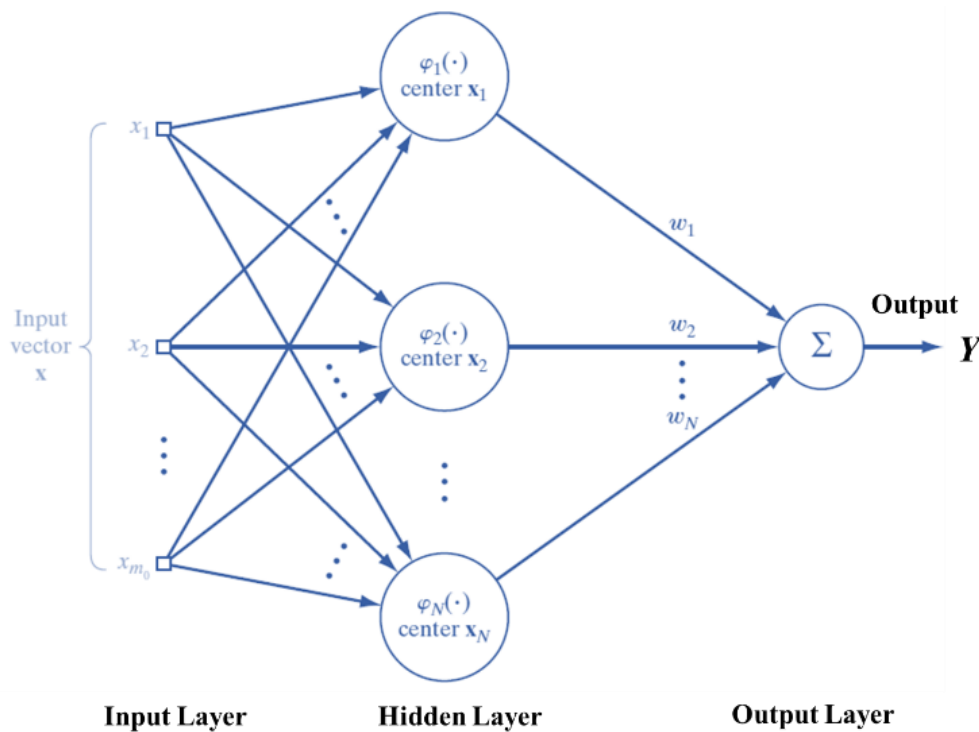


Figure 1 Structure of Radial Basis Function Neural Network

The hidden layer applies a non-linear transformation from the input space to the hidden space. The output layer applies a linear transformation from the hidden space to the output space. The radial basis functions $\varphi_1, \varphi_2, \dots, \varphi_N$ are known as hidden functions while $\{\varphi_i(x)\}_{i=1}^N$ is called the hidden space. The number of basic functions (N) is typically less than the number of data points available for the input data set. Among several radial basis functions, the most commonly used is the Gaussian, which in its one-dimensional representation takes the following form:

$$\varphi(x, \mu) = e^{-\frac{\|x-\mu\|^2}{2d^2}}$$

where μ is the center of the Gaussian function (mean value of x) and d is the distance (radius) from the center of $\varphi(x, \mu)$, which gives a measure of the spread of the Gaussian curve.

The hidden units use the radial basis function. If a Gaussian function is used, the output of each hidden unit depends on the distance of the input x from the center μ . During the training procedure, the center μ and the spread d are the parameters to be determined (Moody and Darken 1989). It can be deduced from the Gaussian radial function that a hidden unit is more sensitive to data points near the center. This sensitivity can be adjusted by controlling the spread d . It can be observed that the larger the spread, the less sensitive radial basis function to the input data. The number of radial basis functions inside the hidden layer depends on the complexity of the mapping to be modeled and not on the size of the data set, which is the case when utilizing multi-layer perceptron ANN. Moreover, RBNN has the ability to recognize a complex relation between the input and output of the model. This research identifies the relationship between the bands and soil types. RBNN model requires some important parameters to be established before perform the training process, such as the performance goal of 0.0005 and the spread constant of 1.”

2a) Sampling design, size and the inference the sample size (25; 19 training, 6 validation) was quite small for such a large area.

Reply:

The authors completely agree with the reviewer that the data used to develop the ANN model is somewhat small in terms of the number of collected data, which may not be recommended for the development of an ANN model. However, the main objective of this research is to propose a methodology for the recognition of soil textures and validate it using data collected on-site. The authors believe that the changes in the soil texture at certain locations within the study area are expected to be relatively minor. More data would be more helpful at different sites of the study area. Therefore, the authors suggest to change the title of this manuscript to reflect the scope of this study, i.e. "Towards the Development of a Spatial Hydrologic Soil Map Using Spectral Reflectance Band Recognition and a Multiple-Output Artificial Neural Network Model".

2b) The chosen smaller area for sampling also did not appear to represent the greater study area (flat accessible area versus a plateau with a dense network of valleys and canyons).

Reply:

The study area (Wadi Horan) is a part of West desert of Iraq, therefore the authors believe that the changes in soil texture at certain areas are expected to be relatively minor. As mentioned in the manuscript, Wadi Horan is flat, and the average topographic incline from east to west is 5 m/km. The plateau with a dense network of valleys and canyons is very small as compared to the overall study area. Therefore, the authors decided to clarify this characteristic of Wadi Horan in the revised manuscript to avoid any confusion.

2c) The validation set represented a very small range of both soil texture separates (sand, silt, clay) and only fell within one of USGS hydrologic group. This limits the inference space to just that group and makes any claims about predicting the other groups correctly unsupported by the data

and result. This makes extrapolation from the smaller sample area to the greater study area unsubstantiated.

- [Reply:](#)

The authors completely agree with the reviewer that this study considered all soil types in the USGS hydrological group during the training stage of model, i.e. as input data shown in Table 1. However, the validation process did not indicate all types of soil in the USGS hydrological soil group even though the regression model has the ability to extrapolate it during the training stage. In this study, the validation exercise resulted in certain types of soil but it does not mean that model is unable to predict other soil types. The model simulation shows that all hydrological soil group types were recognized in the study area, as shown in figure 8.

3a) Overall grasp of literature review and breadth of topics covered and utilized suggest that authors should consider expanding the tools and data they use for making these predictions (beyond having a better sample). For example, in the broad body of digital soil mapping studies, topographical layers from DEM, climate surfaces, and other spectral data (e.g. gamma radiometrics) are often the most effective predictors. I think including DEM variables could prove very effective at this scale, yet this was not done.

[Reply:](#)

The authors agreed with the reviewer that including DEM variables could be very effective at this scale. Also, climate surface, and other spectral data are the most effective predictors. The authors considered the topographical layers from DEM, the Normalized Difference Vegetation Index (NDVI) in dry and wet season, and climate data as shown in line 128. These data extracting techniques are well known and commonly used in research (Cvar 2014; Sayl et al. 2016). The following paragraphs will be added to give more information on the tools and data used in making these predictions (beyond having a better sample):

“The DEM generated from Shutter Radar Topographic Mission SRTM data was used to present the topography of the study area (as shown in Figure 2 below), slope data and drainage. The earth features, such as the Normalized Difference Vegetation Index (NDVI) is used primarily for vegetation identification and to determine the lushness of vegetated land surfaces (using ERDAS Imagine software). These features were considered in this study as a basic guide during the field work. In addition, the location of the points used to generate digital soil map were selected according to DEM generated for a small area in order to overcome the problem of scale”.

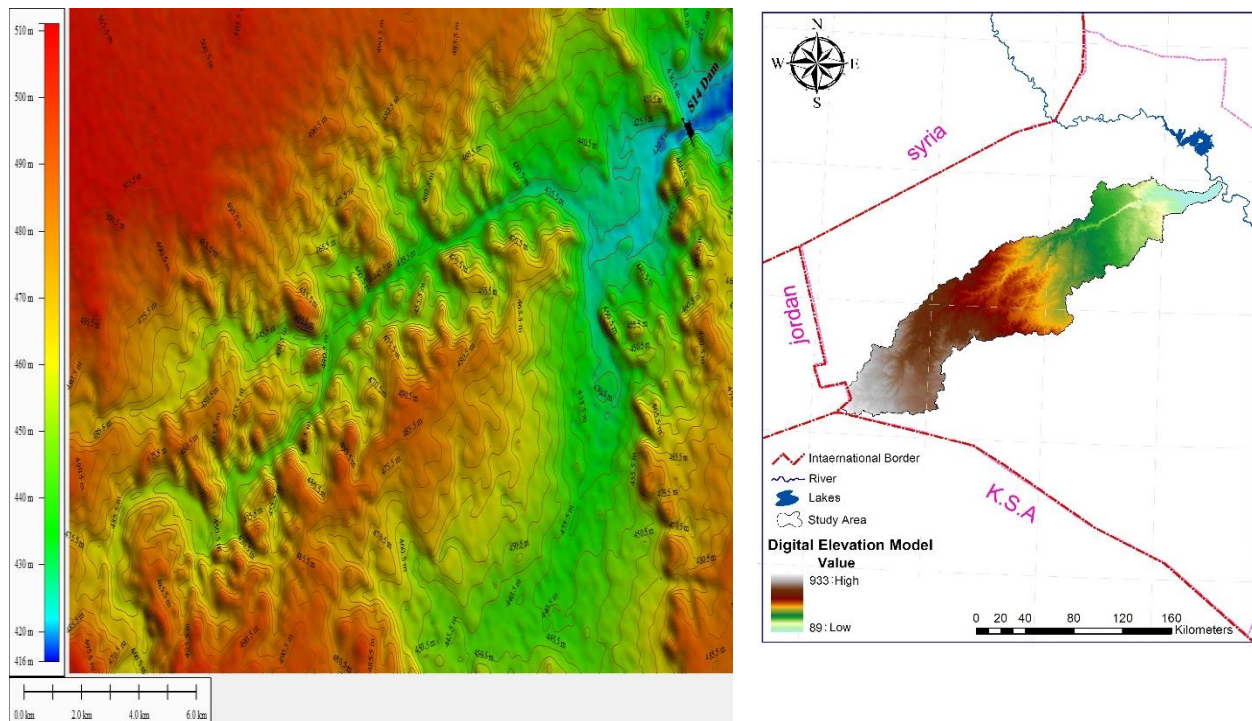


Figure 2 Topography of the location used to generate digital soil map and DEM of the study area generated from Shutter Radar Topographic Mission data

3b) I am also not sure why this form of Neural Network (NNs) was utilized. Generally random forests have been outperforming NNs, why not try other algorithms. If hydro group is the desired target variable, why not predict that parameter? Machine learning seems to do better at classification (particularly random forests).

- [Reply:](#)

The authors would like to highlight that this study is the first attempt to develop an Artificial Intelligence (AI) model for hydrological soil group. There are several AI methods characterized by the efficiency in prediction that can be utilized for random forests and this may be considered in future research. However, the Radial basis neural network (RBNN) was chosen in this study because this method simple and readily available in the Matlab tool box. Additionally, RBNN proved its ability to provide multiple output with a good accuracy in this study.

In this study, the regression model could be further generalized as compared to the classification model. This model built a relationship between the reflectance bands with soil percentages and it is very sensitive. During classification process, the model in training must include all classes and due to the limitation of data set in this study, it is recommended to use regression model than classification. Therefore, this model will be more efficient to find other types of soil based on the output percentages of soil type even if it is not included in training set. Hence, this model setup is intended to be more flexible.

Specific comments

1) Line 20-24 p.1- Hard to tell what was actually mapped? Texture class? Clay%?

- [Reply](#)

The main objective of this study is to determine the distribution of hydrological soil groups in the study area. This is based on the percentages of sand, clay and silt which determined the soil texture. The procedure is given in detail in the Methodology section.

2) Line 31 p. 1- Is the main application this methodology is being created for rainwater harvesting?

- [Reply](#)

Information on hydrological soil group is very important in many areas, such as water resources management and planning, agriculture and other engineering projects. However, the main application for this study is the planning of rainwater harvesting.

3) Line 36 p.2- Please provide citation.

- [Reply](#)

(Melesse and Shih 2002).

4) Line 43 p. 2- It is not an alternative to getting a soil type from field and lab data, it is a way to better utilize the data you have. Any remote sensing model always need actual field and usually lab data to train it.

- [Reply](#)

The authors agree with reviewer that any remote sensing model always need actual field and usually lab data. Therefore, this statement should be replaced by “Remote sensing (RS) represents one of the best solution that may offer possibilities for extending existing soil survey data sets”

5) Line 46 p.2- More appropriate to say “RS can be used to help.

- [Reply](#)

Done.

6) Line 55 p.2- Are you actually talking about Remote Sensing here? The paragraph is confusing.

- [Reply](#)

The authors completely agreed with the reviewer that the paragraph is confusing therefore, the paragraph should be replaced by "Most reported studies revealed high potential of proximal sensing to estimate soil properties based on clear absorption features at the laboratory and local scale. However, for large scale mapping, this exercise need to be extended beyond the plot scale. Important qualitative and quantitative soil information can be obtained from remote sensing data.

Several soil attributes can be determined by spectral analysis under laboratory conditions. However, space borne or airborne spectroscopy complicates the measurements owing to atmospheric influences, lower spectral and spatial resolution, structural effects, spectral mixture of features and geometric distortion (Richter and Schläpfer 2002).

7) Line 58 p.2 – Can be should be used rather than is.

- [Reply](#)

Done.

8) Line 58 p.2 – Five specific should be deleted.

- [Reply](#)

Done.

9) Line 65 p.2 – What is a large area?

- [Reply](#)

The authors will replace ‘large area’ to ‘large scale’.

10) Line 67 p.2 – What correlation? Need to be more specific, and what poor correlation are you referring to?

- [Reply](#)

The authors thank the reviewer for this valuable comment. This line is typo error. The authors have decided to delete this line to avoid any confusion regarding the correlation index.

11) Line 90 p.3 – What constitutes a large area? Must be specific about scale?

- [Reply](#)

The authors will replace ‘large area’ to ‘large scale’.

12) Line 92 p.3 – Combine the first three paragraphs of this section into one- vey wordy and broken as written here.

- [Reply](#)

Done.

13) Line 94-97 p.3 – delete

- [Reply](#)

Done.

14) Line 107 p.4 – Please be more clear here. Are you saying that the shorter valleys are canyons while the longer ones are valleys? How are you defining a canyon? It sounds more like you are trying to describe the variation in the lengths of valleys overall?

- [Reply](#)

The authors thank the reviewer for this valuable comment. This is a typo error. The authors have decided to delete this line to avoid any confusion regarding the study area and replace with following paragraph:

“The main landscape is a plateau that is divided by Wadi Horan, some of it is canyon- like with a few tens of kilometer long, and others are few hundred kilometers in length drain into Wadi Horan.”

15) Line 110 p.4 – Please be more specific, rocky soils? Lots of bedrock outcrops.

- [Reply](#)

The major plateau of the study is rocky soil.

16) Line 111 p.4 – What is structure?

- [Reply](#)

The dip of the strata is almost horizontal, reaching 1° to 2°. The gentle plain reflects the structural position of the study area within the Stable Shelf (Sadooni 1996).

17) Line 113 p.4 – Do you mean deeper soil or proportion of ground soil?

[Reply](#)

It means proportion of ground soil.

18) Line 117 p.4 – This seems out of place, from the earlier mention of water harvesting, it seems that this is the target application for your method. More explanation is needed to related the two and talk more about water harvesting.

- [Reply](#)

The authors completely agreed with the reviewer that more explanation about water harvesting will be added to the revised manuscript “West desert of Iraq is one of the biggest arid regions that has suffered from a severe water shortage, due to its climatic condition and lack of water resources planning and management. When the data are limited or of low quality, decisions related to the planning of rainwater harvesting structures, particularly in the developing countries become more difficult to be made. The nature of most arid regions is generally characterized by the lack of precipitation, high temperature and evaporation, as well as limited surface water and groundwater resources. A rainwater harvesting structure is considered as one of the best solutions to conserve this precious natural resource in the area which has direct effects on both socio-economic development and ecosystem health”.

19) Line 119 p.4 – Combine with next paragraph.

- [Reply](#)

Done.

20) Line 129 p.4 – Also need a subset map showing the location of the sampling area relative to the larger study area?

- [Reply](#)

[Please refer to Figure 3](#)

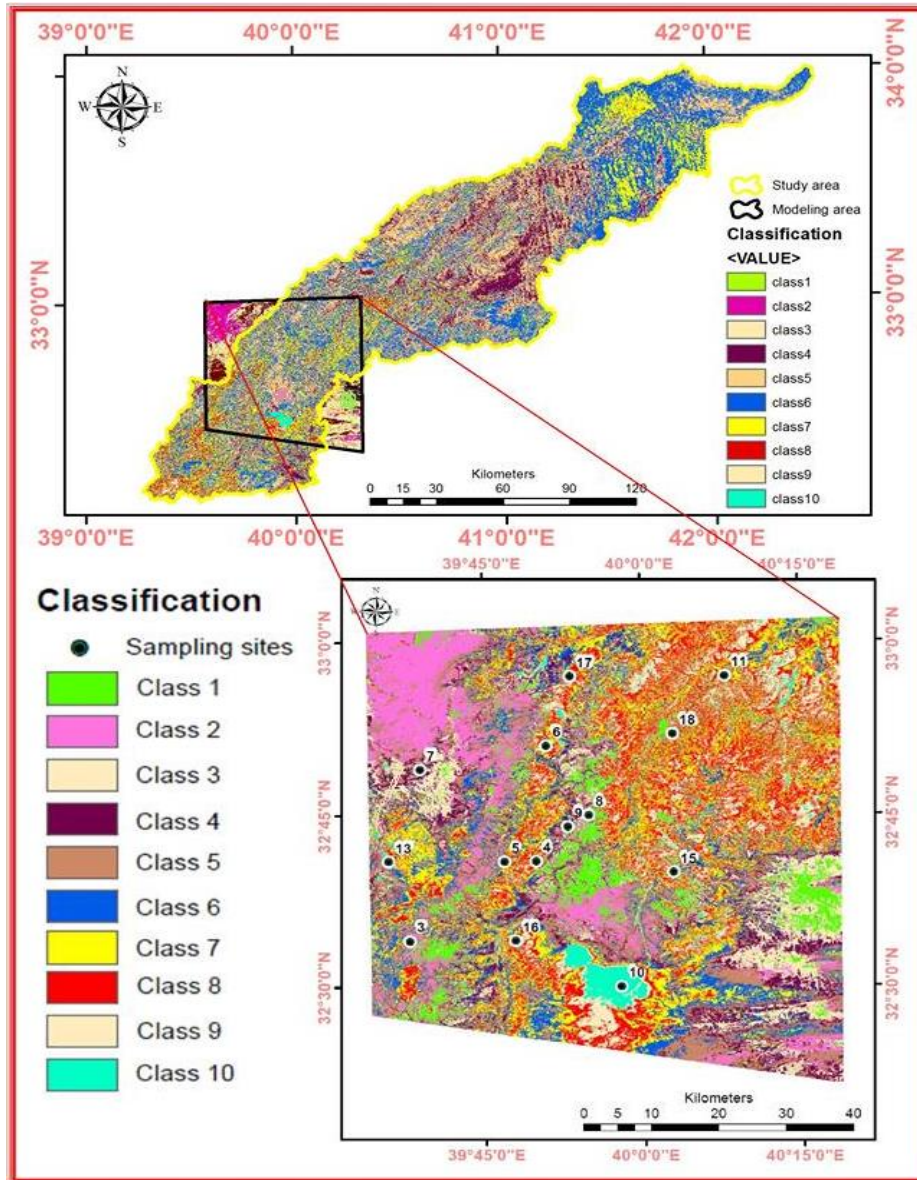


Figure 3 Location of sampling areas

21) Line 130 p.4 – How it is useful?

- [Reply](#)

The primitive map provides a good depiction of some spectral classes and categorized these classes based on the ranges of the image value. This depiction is useful to determine and distribute different kinds of soil properties in this study, which reduces time, labor and cost in the initial stage.

22) Line 140 p.4 – What criteria? Methods must be repeatable.

- [Reply](#)

With reference to line 131, the criteria considered for sampling locations are the error in pixel vegetation cover, the spectral signatures urban areas, water, slope, roads, soil roughness, location and topography.

23) Line 156 p.5 – How were manipulated?

- [Reply](#)

The following paragraphs will be added to give more information on manipulation:

" Arc GIS spatial analyst extension is able to convert the themes, depending on vector features to grids (Huisman and Deby 2009). Additionally, grids can be derived and viewed from various spatial analysis operations. These grid cells have been classified in various ways and different colors were chosen for each class, where they represent the progression of values for a specified data attribute. It is achieved after the raster themes are converted into a shape file, which includes the environmental characteristics that represents the hydrological soil group"

24) Line 161 p.5 – There were never specified any were.

- [Reply](#)

With reference to line 131, the criteria that was consider for sampling location are the error in pixel vegetation cover, the spectral signatures urban areas, water, slope, roads, soil roughness, location and topography.

25) Line 163 p.5 – better than what, based on what evidence? How would looking at that classification tell me more than different visualization of landsat. I can tell a lot about desert lithology and soils by looking at the landsat images.

- [Reply](#)

The word ‘better’ will be deleted. The authors meant that using unsupervised classification is able to reduce the error with spectral signature.

The unsupervised classification is used to classify the physical characteristic and give us wide range of classes than that described through visualization.

26) Line 168 p.5 – To be honest, this whole paragraph and figure 5 are not explained well here, are not document in the methods section, and make very little sense to me.

- [Reply](#)

The sensitivity analysis is a common practice to examine the relationship between input and output parameters and to describe the complicity of this relation. Previous studies such as (Chang and Islam 2000; Proctor et al. 2000; Apan et al. 2002) referred to a relationship between each band with special characteristics of soil and used the sensitivity analysis to validate it. Authors suggest to edit this paragraph to give more information about sensitive analysis. The new paragraph is given below:

“Sensitivity analysis is a prerequisite to determine the reliability of the model through assessment of uncertainties in the output result (Crosetto and Tarantola 2001). Sensitivity analysis is crucial to test the robustness of input and the extent of output variation when parameters are systematic. For this study, a sensitivity analysis was carried out to validate the relationship between soil type and spectral reflectance, as shown in Fig. 5. Soil type could not be detected by band 2 (wavelength (0.45–0.51) μm). Band 9 (1.36–1.38 μm) and band 7 (2.11–2.29 μm) were the most sensitive to soil type, particularly silt and sand, whereas clayey soil could be detected by band 6 (1.57–1.65 μm), band 1 (0.43–0.45 μm) and band 7. Unfortunately, the spectral reflectance for each range of wavelengths represented by the number of bands has a complex relationship with soil type because all these bands participate in detecting the soil texture, but in different weights because of the mineral content of that soil. Because of the variation in spectral reflectance over bands, a highly

accurate model for the estimation of soil type is needed. Therefore, it is important to include all effective bands in the ANN model”.

27) Line 180 p.5 – No clue what you are describing here?

- [Reply](#)

The neural network in this study has multi output where the summation of predicted output equals to the summation of observed data. Therefore, the overall performance was constant.

28) Line 185 p.5 – Not types, these are 3 properties of the soil?

- [Reply](#)

The authors thank the reviewer and the word ‘types’ will be changed to ‘three properties of the soil’.
