A Genetic Programming Approach to Estimate Vegetation Cover in the Context of Soil Erosion Assessment

Cesar Puente, Gustavo Olague, Stephen V. Smith, Stephen H. Bullock, Alejandro Hinojosa-Corona, and Miguel A. González-Botello

Abstract

This work describes a genetic programming (GP) approach that creates vegetation indices (VI's) to automatically detect the sum of healthy, dry, and dead vegetation. Nowadays, it is acknowledged that VI's are the most popular method for extracting vegetation information from satellite imagery. In particular, erosion models like the "Revised Universal Soil Loss Equation" (RUSLE) can use VI's as input to measure the effects of the RUSLE soil cover factor (C). However, the results are generally incomplete, because most indices recognize only healthy vegetation. The aim of this study is to devise a novel approach for designing new VI's that are better - correlated with C, using field and satellite information. Our approach consists on stating the problem in terms of optimization through GP learning, building novel indices by iteratively recombining a set of numerical operators and spectral channels until the best composite operator is found. Experimental results illustrate the efficiency and reliability of our approach in contrast with traditional indices like those of the NDVI and SAVI family. This study provides evidence that similar problems related to soil erosion assessment could be analyzed with our proposed methodology.

Introduction

Soil erosion is a complex phenomenon that detaches and transports soil material through the action of an erosive agent. Soil erosion by flowing water on slopes is an important land degradation problem at a global scale, because it is strongly influenced by human activities such as agriculture and generates major environmental impacts and high economic costs. Most estimates of soil erosion, as undertaken by agricultural scientists, are at field scales (Brady and Weil, 2008). The aim of erosion research at regional scales is a general evaluation of the landscape and its susceptibility to soil erosion, taking into account only the main factors influencing the process. Therefore, the erosion assessment at regional scale is usually based on empirical models or expert evaluation. The most widely applied models are the Universal Soil Loss Equation (USLE) (Wischmeier and Smith, 1978), and its revision, the Revised Universal Soil Loss Equation (RUSLE) (Renard *et al.*, 1997). USLE and RUSLE are statistically-based water erosion models related to six erosional factors:

$$A = R * K * L * S * C * P, \tag{1}$$

where R and K set the dimensions of A as the average of soil loss in Mg ha^{-1} yr^{-1}. R is the rainfall-runoff factor and is measured as the product of total storm energy, E, and the maximum 30 minute intensity, I_{30}, for all storms in a year, so its units are described in terms of MJ mm ha^{-1} h^{-1} yr^{-1}. On the other hand, K represents the influence of soil properties related to soil texture and structure on soil loss during storm events; the units of K are Mg h MJ^{-1} mm^{-1}. The remaining factors are dimensionless and serve to scale erosion relative to standard experimental conditions, which are described within the USLE and RUSLE manuals cited above. These scaling values range from 0 meaning no erosion, to numbers greater than 1, where erosion is more rapid than the experimental conditions. In this way, L represents the slope length factor, S is the slope steepness factor, C provides the ground cover factor, and P describes the conservation support practice factor.

From the standpoint of soil conservation planning, the C factor is one of the most important parameters of RUSLE, because it measures the combined effect of all interrelated cover and management variables. Vegetation cover acts as a protective layer between the atmospheric elements and soil. For example, live or dead leaves and stems absorb most of the energy of raindrops and surface water to decrease the volume of rain reaching the soil surface (Asis and Omasa,

Cesar Puente is with the Departamento de Geociencias Ambientales, EvoVisión Project,Centro de Investigación Cientifica y de Educación Superior de Ensenada (CICESE), Carretera Ensenada-Tijuana 3918 Zona Playitas, 22860, Ensenada, Baja California, México (cpuente@cicese.mx).

Gustavo Olague is with EvoVisión Project, Centro de Investigación Cientifica y de Educación Superior de Ensenada (CICESE).

Stephen V. Smith and Alejandro Hinojosa-Corona are with the Departamento de Geología, Centro de Investigación Cientifica y de Educación Superior de Ensenada (CICESE).

Stephen H. Bullock is with the Departamento de Biología de la Conservación, Centro de Investigación Cientifica y de Educación Superior de Ensenada (CICESE).

Miguel A. González-Botello is with the Maestría en manejo de ecosistemas de zonas áridas, Facultad de Ciencias, Universidad Autónoma de Baja Californa (UABC).

Photogrammetric Engineering & Remote Sensing
Vol. 77, No. 4, April 2011, pp. 363–376.

0099-1112/11/7704–0363/$3.00/0

2007). Roots also contribute to the mechanical strength of soil. While the C factor is readily measured at local plot scales for individual crops grown under standard conditions, it is difficult to quantify over broad geographic areas of mixed vegetation cover due to the labor and time required for large numbers of individual measurements. In order to avoid such problems, methods for extracting ground cover information from satellite imagery have become powerful tools to estimate biophysical properties from plants and other objects protecting the soil surface (Ryerson, 1999).

There are three main approaches to the problem of extracting C from satellite imagery as tools to generalize local field plot samples to a broad area. The cover classification method, the Linear Spectral Mixture Analysis, and the Vegetation Index method. Traditionally, this task has been achieved through the cover classification method, which consists of assigning C values to labels that correspond to specific characteristics of the Earth's surface (Ryerson, 1999; Folly *et al.*, 1996). This method, however, results in C factor estimates that are homogeneous for relatively large areas, and do not adequately reflect the spatial variation in vegetation density within cover classes over large geographic areas (Wang *et al.*, 2002). Another approach for estimating vegetation fractional cover is known as Linear Spectral Mixture Analysis (LSMA) that has proven useful for detecting photosynthetic and non-photosynthetic vegetation, as well as bare soil (e.g., Asner and Lobell, 2000; Guerschman *et al.*, 2009). In particular, Asis and Omasa (2007) proposed a technique based on LSMA in order to estimate the C factor as a function of the vegetation fractional cover. This approach exploits the capacity of LSMA to estimate the fractional abundance of ground vegetation and bare soil simultaneously in one pixel. Hence, the classification is at a sub-pixel level, giving more variability to C map. However, in order to perform an efficient un-mixing process that avoids misclassification, it is necessary to know *a priori* the components being measured within a pixel.

In the literature several vegetation indices (VI's) had been explored to determine C. A vegetation index (VI) applies a mathematical formula to the spectral channels of satellite imagery in order to enhance the signal representing the vegetation cover. Thus, VI's are correlated with the C factor using regression analysis (mainly linear regression). However, several studies reported low correlations between available empirical VI's and C (e.g., Asis and Omasa, 2007; de Jong, 1994; Smith *et al.*, 2007). One reason is that the response of the VI's are focused mainly on healthy (green) vegetation, and not on senescent (dry and dead) vegetation, which can also be an important contributor to C.

Problem Statement

The previous discussion shows that the information of vegetation required by erosion models differs from the information provided by conventional VI's. According to de Jong (1994), the evaluation of soil protective cover with VI's gives a good correlation as long as the vegetation is green, but gives less satisfactory results for senescent vegetation. Hence, for the erosion process the condition of the vegetation is of minor importance because senescent vegetation will protect the soil as well as vigorous vegetation. Nevertheless, most of the VI's used to estimate the RUSLE cover factor were designed to distinguish healthy vegetation because those indices use ρ_{red} and ρ_{NIR} bands only. Moreover, the combinatorial possibilities of using other spectral channels have not been comprehensively explored. A major difference with previous works is that such indices have been designed by experts using traditional representations that have a clear and preferably physical well-founded definition. In this paper, we will show how a powerful machine learning strategy known as genetic programming (GP) is able to create synthetic VI's that can outperform previous man-made or manually-designed VI's in estimating the RUSLE soil cover factor. For this task the problem is posed as a search problem, where the objective is to find the index that correlates best with C factor data. The GP algorithm then builds new indices by iteratively recombining a set of numerical operators and channels until the best composite index is found. This paper follows a line of inquiry of previous studies, in which vegetation indices were correlated with C (e.g., Asis and Omasa, 2007; de Jong, 1994; Smith *et al.*, 2007). We undertook a field survey in order to derive the vegetation indices that best describe C for an area in northwestern Baja California, Mexico (see Figure 1). This study is the first step towards modeling regional soil erosion caused by water using genetic programming (GP) as a combinatorial engine that synthesizes indices that better correlate with on-site information than conventional indices taken from the literature. In particular, this paper focuses on the cover factor of the RUSLE model using our GP approach.

Brief Introduction to GP

Genetic programming (GP) can be defined as an evolutionary computation technique inspired from the principles of biological evolution (Poli *et al.*, 2008). GP is an offshoot of genetic algorithms, and is able to evolve a population of computer programs (mathematical expressions or formulae) that learn a user-defined function. GP starts with an initial population of randomly generated programs. Each individual computer program in the population is measured in terms of how well it performs in the particular problem environment. The computer programs in the initial population of the process will generally have poor fitness. Nonetheless, some individuals in the population will turn out to be somewhat more fit than others. The Darwinian principle of reproduction and survival of the fittest, as well as the genetic operators of crossover and mutation are used to create a new offspring

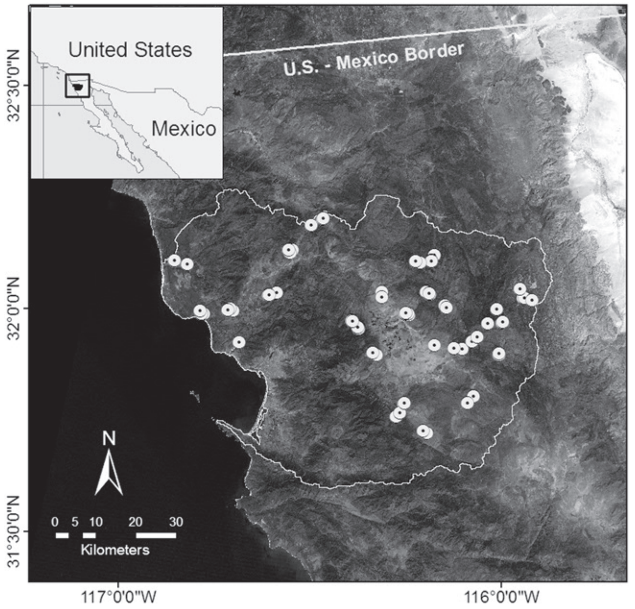


Figure 1. Map of Baja California, Mexico showing the geographical location of Todos Santos Watershed. The figure also shows the location of field samples used to estimate the C factor.

population of programs. Thus, generation by generation, GP stochastically transforms populations of programs into new, generally better, populations of programs. As is true for any other stochastic process, the GP analysis can never guarantee results; however, its heuristical search strategy can lead it to escape traps that challenge deterministic methods.

Evolutionary computation techniques have been successfully applied to photogrammetric problems, and multispectral analysis and remote sensing problems, such as edge detection, image segmentation and classification, and feature extraction. For instance in photogrammetry, Olague (2002) describes the use of genetic algorithms (GA) for automating the photogrammetric network design process. In feature extraction, Bhanu and Lin (2003) use a genetic algorithm to select a set of features to discriminate desired targets from the natural clutter false alarms in SAR (Synthetic Aperture Radar) images. Furthermore, Daida *et al.* (1996) devised a GP approach to identify ice-flow ridges from SAR imagery. In the same way, the work made by Howard *et al.* (1999) automates ship detection from SAR images of the English Channel taken by the European Remote Sensing (ERS) satellite using a two-stage GP process. Also, Ross *et al.* (2005) used GP to evolve Boolean and general mathematical expressions in order to discriminate among three specific minerals (buddingtonite, alunite, and kaolinite) from hyperspectral images. With respect to classification and segmentation, Rauss *et al.* (2000) evolved genetic programs to classify hyperspectral imagery. Harvey *et al.* (2002) devised GENIE (GENetic Imagery Exploitation) system, which uses a hybrid combination of linear genetic programming with conventional classifier algorithms to cope with hyperspectral image classification. Recently, Makkeasorn *et al.* (2009) developed a two-stage GP algorithm to seek for the best vegetation index in conjunction with soil moisture variation to aid in the classification of riparian zones in a semi-arid landscape. The main contributions of our work are the following:

1. The methodology describes for the first time a way of synthesizing VI's using field data and satellite information.
2. As a consequence of the proposed methodology, it is possible to identify the basic components responsible of good VI's, which are specifically designed to solve a particular problem.

Thus, we take advantage of these contributions by obtaining a reliable estimate of the C factor, according to field measurements based on the RUSLE protocol (Dissmeyer and Foster, 1980; Renard *et al.*, 1997; Weltz *et al.*, 1987; Wischmeier and Smith, 1978). The structure of this paper is as follows. The following section presents the materials and methods used to develop our study. We begin with a description of the study area in order to state the physical conditions of our work. Then, we give an outline of the preparation of field data and satellite imagery, and how to relate vegetation indices to C. A brief description of the Genetic Programming approach and a complete explanation of our methodology are also presented in this section. The next Section provides experimental results in order to illustrate the quality of our synthetic vegetation indices followed by our conclusions.

Materials and Methods

Study Area

The study area is the Todos Santos watershed located in northwestern Baja California, Mexico between 31°35′ to 32°16′N latitude and 116°53′ to 115°51′W longitude and is defined and described by Smith *et al.* (2007) (see Figure 1). The watershed covers an area of approximately 4,900 km^2; the climate is Mediterranean, with cool, moist winters and warm, dry summers. The area is topographically complex, elevation ranges between sea level and 1,876 m. It is characterized by large inter-mountain alluvial valleys, alluvial deposits in stream-beds, and an alluvial coastal plain. Much of the coastal plain and middle-elevation alluvial area have been cleared for urban and agricultural development. However, coastal scrub occurs along a belt of approximately 50 km inland and/or 500 m elevation. Another type of scrubland, Chaparral (evergreen *sclerophyllous* shrubs), covers the upper slopes of coastal and inland hills. A plateau at the upper limit of the watershed hosts pine forest and mountain meadows. Hence, the land-cover of the watershed can be summarized as follows: Shrub-land (chaparral and coastal scrub) occupy approximately 69 percent of the total watershed, while agricultural and grazing areas occupy about 22 percent. Other cover (evergreen, mixed woodlands, urban areas, etc.) occupies 9 percent.

Field Data Collection

The fieldwork involved the measurement of different parameters for the C factor estimation. Experimentally, the C factor would be evaluated from long-term experiments where soil loss would be measured from field plots for which the other factors in Equation 1 were known. However, in the absence of long-term experimental data and to have estimates for any variety of site conditions, it is possible to estimate the C factor by using standard calibration of sub-factors. Wischmeier and Smith (1978) identified three major sub-factors that determine the effectiveness of vegetation in limiting soil erosion on rangelands. First, the soil surface cover sub-factor is related to the fractional cover of the soil surface by non-eroding material (basal area of plants, rocks, and organic litter). Second, the canopy cover sub-factor is related to the fractional cover of the soil surface provided by above-ground plant biomass and the height that raindrops fall after leaving the plant to impact the soil surface. Third, the residual and tillage sub-factor is based on the effects made of root biomass, other organic matter in the soil, compaction, and surface stabilization.

Prior to the fieldwork, a detailed examination of satellite imagery and topographic maps of the watershed was conducted in order to identify representative sampling sites. Site selection was subject to local uniformity regarding NDVI and elevation/slope over at least a hectare surrounding the specific measurement site. Hence, field sampling was conducted between February and May 2007. A total of sixty-seven (67) sampling sites were located and established in the study area (see Figure 1; González-Botello and Bullock, unpublished report, 2009). At each site, the percentage of canopy cover was estimated using a line transect sampling method described by Bauer (1943) and Zippin and Vanderwier (1994). Measurements were taken at 20 different random points along the 30 m transects placed at the perpendicular direction to the predominant slope.

At each point, a plumb was dropped, then the surface cover percentage was visually estimated within a 10 cm diameter micro-plot around the plumb. Each observation was classified using five linear categories in accordance with its cover percentage (0 = 0 to 1%, 1 = 1 to 25%, 2 = 25 to 50%, 3 = 50 to 75%, and 4 = 75 to 100%). Hence, in order to assign a unique value of surface cover (g) for each transect, the average category of the 20 points was obtained. Similarly, the height above ground of the lowest plant structure touching the plumb line was registered (drop height), and the values were averaged within each transect. Thus, the height (h) and percentage of canopy cover (p) was established. Finally, the surface roughness (r) was assessed by adjusting field values to Table 5 and Table 6 from the RUSLE model description (Renard *et al.*, 1997) using an empirical scale. Buried root biomass (b) was defined by primary productivity according to methodology described by Weltz *et al.*, (1987), which was assumed to be uniform as a long-term average. Hence, the C factor was estimated for each sampling site by using the following equation derived from Table 10 in Wishmeier and Smith (1978):

$$C = 0.45(e^{[-0.012 \cdot b]}) \cdot (1-p \cdot e^{[-0.328 \cdot h]}) \cdot e^{(-0.039 \cdot g \cdot [0.24/r]\hat{}0.08)} \quad (2)$$

This equation is similar to equations described by Weltz *et al.* (1987) and Renard *et al.* (1997).

Satellite Imagery Preprocessing

A Landsat-5 TM scene (path 39, rows 037 and 038), taken on 13 April 2007 was downloaded from USGS[1] and analyzed for this study. Landsat TM has six visible/near infrared bands (0.48 to 2.3 μm) with a spatial resolution of 30 m, and one thermal band (11μm) with a spatial resolution of 60 m. However the thermal band was not used in this work. We used a Dark Pixel Correction (DPC) formula (Vincent, 1997) to perform atmospheric effect correction on each band of the TM image dataset. Then, the digital number (DN) of TM bands 1 to 5 and 7 were converted to exo-atmospheric reflectance units as described by Chander and Markham (2003). Also the satellite imagery was geometrically rectified using topographic maps and well-known ground control points in order to accurately link it to ground data. Image data were extracted and matched with field data as follows. In each image, the 67 field sites were located. A window of 3-by-3 pixels was selected around each sampling site, according to the fact that each sampling site had been chosen to be uniformly covered over at least a hectare (100-by-100 meters). Thus, for each window of nine pixels, the median value was extracted and labeled with the corresponding field site. Then, the whole data set was divided into two parts. A set of 47 points, named "training dataset", and a set of 20 points, named "testing dataset." The aim of both datasets will be elaborated in the methodology subsection.

Relating the C factor to Vegetation Indices

Vegetation indices (VI's) typically quantify the vigor of vegetated land-cover, for example, by estimating vegetation greenness. A vegetation index (VI) applies a mathematical formula to the spectral channels of satellite imagery, generally ratios or differences among bands, principal component analysis, or other linear or non-linear combinations of bands in order to enhance the signal of the vegetation cover presented on the surface of the Earth.

TABLE 1. CONVENTIONAL VEGETATION INDICES AND THEIR CORRELATION COEFFICIENT ($r_{x,\,y}$)

Index	Description	$\lvert r_{train} \rvert$
RVI4	$\rho_{\mathrm{SWIR1}}/\rho_{\mathrm{SWIR2}}$	0.512
RVI5	$\rho_{\mathrm{SWIR1}}/\rho_{\mathrm{red}}$	0.383
RVI1	$\rho_{\mathrm{NIR}}/\rho_{\mathrm{red}}$	0.342
GEMI	$\eta(1 - 0.25\eta) - (\rho_{\mathrm{red}} - 0.125)/(1 - \rho_{\mathrm{red}})$ where $\eta = [2(\rho_{\mathrm{NIR}}{}^2 - \rho_{\mathrm{red}}{}^2) +$ $1.5\rho_{\mathrm{NIR}} + 0.5\rho_{\mathrm{red}}]/(\rho_{\mathrm{NIR}} + \rho_{\mathrm{red}} + 0.5)$	0.329
RVI2	$\rho_{\mathrm{NIR}}/\rho_{\mathrm{green}}$	0.329
NDVI	$(\rho_{\mathrm{NIR}} - \rho_{\mathrm{red}})/(\rho_{\mathrm{NIR}} + \rho_{\mathrm{red}})$	0.314
SAVI2	$\rho_{\mathrm{NIR}}/(\rho_{\mathrm{red}} + b/a)$	0.309
IPVI	$\rho_{\mathrm{NIR}}/(\rho_{\mathrm{NIR}} + \rho_{\mathrm{red}})$	0.314
SASI	$\beta_{\mathrm{SWIR1}}(\rho_{\mathrm{NIR}} - \rho_{\mathrm{SWIR2}})$	0.254
GVI	$-0.3344\rho_{\mathrm{blue}} - 0.3544\rho_{\mathrm{green}} -$ $0.4556\rho_{\mathrm{red}} + 0.6966\rho_{\mathrm{NIR}}$ $+ 0.0242\rho_{\mathrm{SWIR1}} - 0.2630\rho_{\mathrm{SWIR2}}$	0.228
EVI	$G((\rho_{\mathrm{NIR}} - \rho_{\mathrm{red}})/(\rho_{\mathrm{NIR}} + C1{*}\rho_{\mathrm{red}} - C2{*}\rho_{\mathrm{blue}} + L))$ where $G = 2.5; C1 = 6; C2 = 7.5; L = 1$	0.101

Vegetation indices have been explored for mapping the C factor by regression analysis (primarily linear regression; e.g., Asis and Omasa, 2007; de Jong, 1994; Sabins, 1997; Smith *et al.*, 2007). However, these studies report low correlation with available vegetation indices. The reason is that the response of conventional vegetation indices is focused on healthy vegetation. Most of these VI's only use ρ_{red} and ρ_{NIR} bands in various combinations, because these are the bands that distinguish healthy vegetation. Examples include the Ratio Vegetation Index (RVI) (Jordan, 1969), the Soil Adjusted Vegetation Index (SAVI) (Huete, 1988) and its improvements, and the well-known Normalized Difference Vegetation Index (NDVI) (Rouse *et al.*, 1973; Tucker, 1979). NDVI and RVI have been widely-used to estimate the C factor.

There are some indices that use additional bands, but none of them have been used to estimate the RUSLE C factor. For instance, Pinty and Verstraete (1992) proposed the Global Environmental Monitoring Index (GEMI), which is designed to minimize atmospheric noise and uses band ρ_{blue}. The Green Vegetation Index (GVI) developed by Crist and Cicone (1984), extends the concept of the soil line in the $\rho_{\mathrm{red}}-\rho_{\mathrm{NIR}}$ spectral space to a soil hyperplane in the space of several bands by constructing a linear combination of these bands (see Table 1). The Normalized Difference Water Index (NDWI) reported by Gao (1996), uses the shortwave infrared region of the spectra (SWIR bands) to detect liquid water, wet soil and plant moisture. Recently, Khana *et al.* (2007) developed a novel approach to parameterize the shape of a part of the spectral signature by measuring the angle formed between three consecutive bands. One of these angle indices is the Shortwave Angle Slope Index (SASI) (Table 1), which is a combination of NIR, SWIR1, and SWIR2 of MODIS sensor bands, and it is useful to detect moisture contents.

In order to define the relation between conventional VI's and C factor field samples, thirty vegetation indices extracted from the literature were reviewed and applied over our imagery data. VI's were related to field data as follows. First, the indices were calculated for each field site from training dataset using the median values previously extracted from satellite data. Thus, the VI's values were combined with the corresponding C factor field values. The combination yielded a matrix of two columns and 47 rows. Then, the correlation coefficient was computed over this matrix. Note that a negative correlation can exist. We use $|r_{x,y}|$ because a solution with $r_{x,y}$ of -1 should be as good as a solution with $r_{x,y}$ of 1. The best ten VI's of those thirty calculated VI's are shown at Table 1. We can appreciate in Table 1 that the VI's with the best performance are variations of the RVI (which in fact, stands at the third place). We decided to test five ratios using different combinations of bands. Four of those ratios are at the top ten, proving that the *ratio structure* is a simple, yet very useful way to construct a vegetation index. Note that the widely-used NDVI and EVI (based on Red and Near-Infrared reflectivities) are far from the best. The best is the so-called RVI4 because it uses short wave infrared reflectivities, which are able to detect plant moisture and lignin/cellulose components, more evident on senescent vegetation. Hence, the low correlation coefficients obtained in general with conventional VI's show that the information of vegetation required by erosion models differs from the information provided by current VI's.

Despite the drawbacks we have noted, VI's have gained increasing popularity in mapping RUSLE's C factor (e.g., Asis and Omasa, 2007; de Jong, 1994; Lin *et al.*, 2002; and Lu *et al.*, 2003; Smith *et al.*, 2007). We believe that VI's represent a challenging search space where the

[1] USGS Global Visualization Viewer at **http://glovis.usgs.gov**

combinatorial possibilities not yet explored by experts makes it a promising research area. In this work a GP approach is chosen to exploit that search space.

Synthesizing Vegetation Indices through GP

GP has been very successful at evolving novel and unexpected ways of solving problems producing a number of instances of results that are competitive with human-produced results (Koza, 2010). The executable nature of the individuals evolved makes GP ideally suited for implementing multispectral analysis and other remote sensing applications, such as classifiers and mathematical indices that enhance the signal of several features over the Earth's surface by means of supervised learning approaches (Agnelli *et al.*, 2002).

The basic steps in a GP system are shown in Table 2. For a graphic overview refer to the GP diagram from Figure 2. GP discovers how well a program works by running it, and then comparing its behavior with some ideal (line 3). We might be interested, as with this case, in how well a program can find the implicit relation between the response of a sensor and the field where that sensor was used. This relation is quantified to give a numeric value called *fitness.* The computer programs in the initial generation of the process will generally have poor fitness. Nonetheless, some individuals in the population will turn out to be somewhat more fit than others. Those programs that do well are chosen for *mating* (line 4) and produce new programs for the next generation (line 5). The primary genetic operations that are used to create new programs from existing ones are *crossover* and *mutation.* The former is defined as the creation of a *child program* by combining randomly chosen parts from two selected *parent programs*; and the latter is the creation of a new child program by randomly altering a randomly chosen part of a selected parent program.

TABLE 2. GENETIC PROGRAMMING ALGORITHM

```
1. Randomly create an initial population of programs from the
   available primitives.
2. Repeat . . .
3. Execute each program and compute its fitness.
4. Select one or two program(s) from the population with a
   probability based on fitness to participate in genetic
   recombination.
5. Create new individual program(s) by applying genetic operations
   with specified probabilities.
6. until an acceptable solution is found or some other stopping
   condition is met (e.g., a maximum number of generations is
   reached).
7. return the best individual up to this point.
```

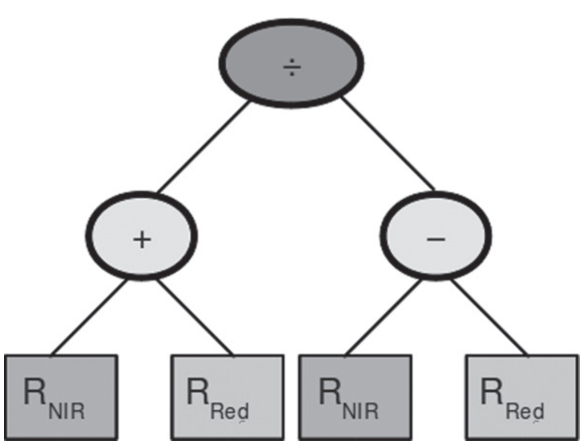


Figure 2. GP syntax tree representing program for NDVI.

In this section, we describe the three major preparatory steps for applying GP to the creation of vegetation indices; first the definition of the Terminals and Functions is set; second, the fitness measure; and third, the parameters for controlling the algorithm and the criterion for terminating a run. Finally, we delineate the process to evaluate our results.

Representation and Search Space

In our work, candidate solutions being evolved by GP process are encoded through tree-based representations that match the mathematical expressions of the vegetation indices. For example Figure 2 shows the tree representation of the NDVI index. The variables and constants in the program (in this case the reflectance values of NIR and Red bands) are leaves of the tree. In the GP process they are called *Terminals*, while the arithmetic operations (+, −, and ÷) are internal nodes, called *Functions.* The sets of allowed Functions and Terminals are defined as the primitive set of a GP system, which represents the problem search space.

The primitive set of our GP system is presented in Table 3. The features that comprise the primitive set were stated as follows. The Terminal set was represented by information on the spectral bands, for example the bands ρ_{red}, ρ_{NIR}, etc., and angles based on that bands like β_{green}, β_{red}, etc. Angles are formed at any vertex in the multispectral broadband spectrum, as defined by Khana *et al.* (2007). For instance β_{green} is a combination of ρ_{blue}, ρ_{green}, and ρ_{red}. The equation for calculating β_{green} is shown below:

$$\beta_{\text{green}} = \cos^{-1}((a^2 + b^2 - c^2)/(2*a*b)), \tag{3}$$

where a, b, and c are Euclidean distances between vertices ρ_{blue} and ρ_{green}; ρ_{green} and ρ_{red}; and ρ_{blue} and ρ_{red}, respectively. The same equation can be applied for every angle. See Khana *et al.* (2007) for further details. In addition, we considered the a and b terminals, which represent the soil line parameters (slope and y-intercept respectively). We determine soil line parameters by manually extracting reflectance characteristics of bare soil pixels from our Landsat imagery, followed by the adjustment of a line to these pixels. Finally, to complete the Terminal set we considered the top ten conventional vegetation indices with the best performance over our study area, represented by I_{RVI4}, I_{GEMI}, etc., described - previously (Table 1). Note that despite the fact that NDVI and EVI are not part of the top ten, we decided to include them as I_{NDVI} and I_{EVI}, because they are broadly-used indices. The function set was represented by arithmetic operations (+, −, and *) because these functions are widely used in common VI's design. Furthermore, we decided to evaluate

TABLE 3. FEATURES THAT INTEGRATE THE PRIMITIVE SET

Features	Description
Terminals	
ρ_{blue}, ρ_{green}, ρ_{red}, ρ_{NIR}, ρ_{SWIR1}, ρ_{SWIR2}	Bands of the Landsat-5 TM multi-spectral image
β_{green}, β_{red}, β_{NIR}, β_{SWIR1}	Angles based on bands (according to Khana *et al.* (2007))
I_{RVI1}, I_{RVI2}, I_{RVI4}, I_{RVI5}, I_{GEMI}, I_{NDVI},I_{EVI}	Best-performance and common conventional indices
a, b	Slope and y-intercept of the soil line
Functions	
+, −, *	Arithmetic operators
NDSI, RSI	Composite operators

explicitly the arithmetic structure of the broadly-used NDVI and RVI by adding two composite operators: the Normalized Difference Spectral Index (NDSI) and the Ratio Spectral Index (RSI), which are defined as follows:

$$\text{NDSI}[i, j] = (\text{R}_i - \text{R}_j)/(\text{R}_i - \text{R}_j), \text{ and} \tag{4}$$

$$\text{RSI}[i, j] = \text{R}_i/\text{R}_j, \tag{5}$$

where R_k represents any reflectance at a single band *k*. NDSI has been used in previous studies. For example, Inoue *et al.* (2008) used NDSI to explore spectral indices for estimating photosynthetic variables from hyperspectral imagery.

Fitness Function.

The fitness function was based on the correlation coefficient $r_{x,y}$ that indicates the strength and direction of the linear relationship between the C factor and each synthesized vegetation index. In this work we choose to apply the absolute value operator of $r_{x,y}$ because the closer the coefficient is to either −1 or 1, the stronger the correlation between the variables. Hence, the fitness function is defined as follows:

$$\text{Q} = \max(|r_{x,y}|), \text{such as } r_{x,y} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \tag{6}$$

$$= \frac{\text{E}((x - \mu_x)(y - \mu_y))}{\sigma_x \sigma_y},$$

where E is the expected value and cov is covariance; *x* represents the RUSLE C factor, *y* is the synthesized vegetation index and $r_{x,y}$ is defined in the range $\{r_{x,y}: -1 \leq r_{x,y} \leq 1\}$.

Initialization, GP Parameters, and Solution Designation.

GP process was programmed in MatLab with the GP toolbox GPLAB (Silva, 2007). Table 4 provides the GP runtime parameters used during the experimental test. We carried out a test-and-error process in our experiment to get the suitable values for some parameters. For others, we used existing settings, which have been reported to work well across a variety of applications. The termination criteria was defined by a maximum number of generations (50 in this case); thus, the evolutionary process reaches an optimum index for each single run. Refer to Koza (1992) and Poli *et al.* (2008) for more details about definition of parameters and settings.

Figure 3 shows the flowchart of the procedure developed to generate novel VI's in order to estimate the C factor. After the image preprocessing step, the initial population of

Table 4. Parameters Used for GP Training

Process configuration	
Generations	50
Population size	50 individuals
Initialization	Ramped Half-and-Half
Crossover probability	0.70
Mutation probability	0.3
Tree depth	Dynamic depth selection
Dynamic max. depth	3 levels
Real max. depth	4 levels
Selection	Lexicographic parsimony pressure tournament
Elitism criteria	Keep the best individual

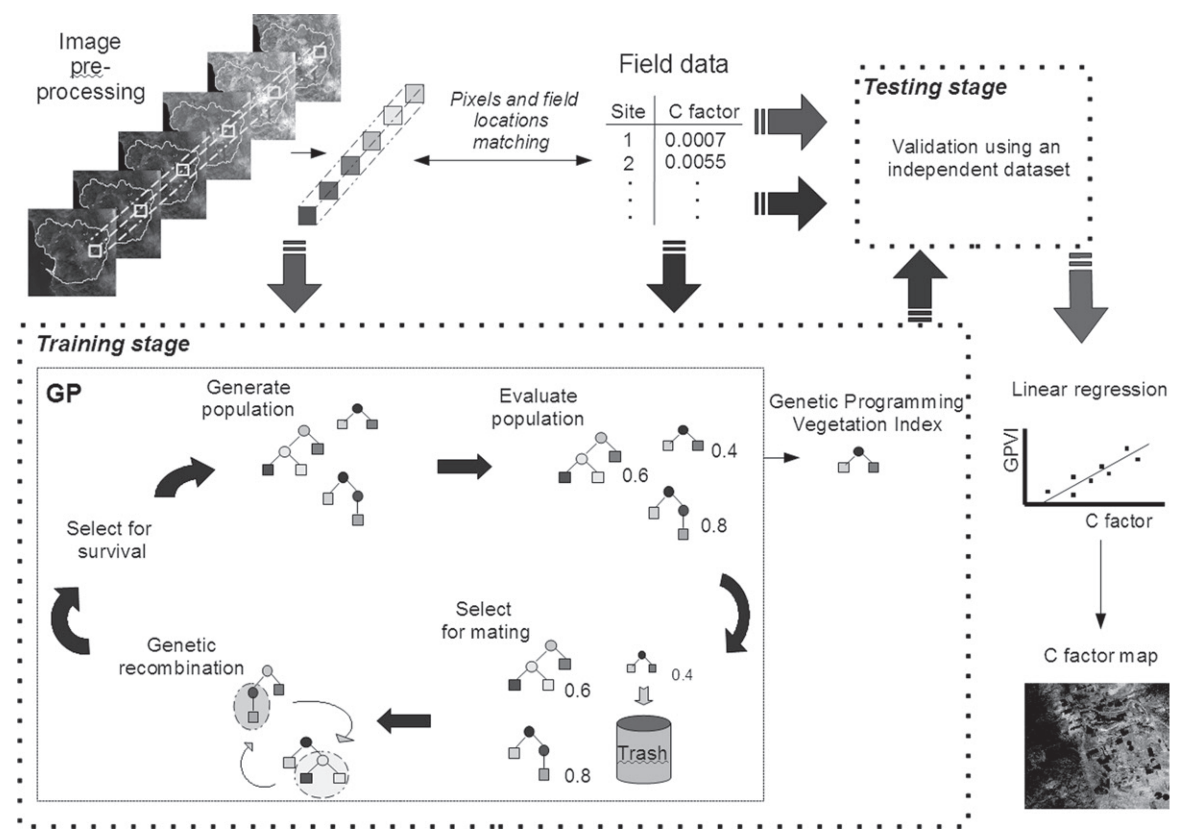


Figure 3. General flowchart of the methodology to estimate C from vegetation indices synthesized by GP.

solutions is generated by randomly combining elements of Terminals and Functions sets. After that, each individual of the initial population is evaluated by the fitness function. The next step is to select candidate solutions in order to rank all individuals and discard the solutions with low fitness. Then, the genetic recombination between selected trees is performed through crossover or mutation operators. Finally, the next generation results from the survival selection of the best parent and child solutions. These steps are iterated until the maximum number of generations is reached (50 generations in our case). The best solution is kept from the population of the 50th generation. This new synthetic vegetation index is called GPVI*j*, where *j* stands for the run number. We will use the acronym GPVI (Genetic Programming Vegetation Index) to indicate our synthesized indices.

Accuracy Evaluation

The accuracy evaluation is implemented as usually done in general machine learning through the two stages of training and testing. The accuracy assessment was performed for each GPVI by computing the absolute difference between correlation coefficient obtained for the training and testing datasets: $D = \lvert r_{\mathrm{train}} - r_{\mathrm{test}} \rvert$.

A way of evaluating the importance of each element at the primitive set is a frequency of use (FOU) unit, which implies pattern recognition capability of the GP process (Makkeasorn *et al.*, 2009). The "survival of the fittest" approach that is implicit in the algorithm, allows GP to extract relevant elements useful to build VI's from the pool of elements. Based on this idea, better quality elements will be used more often than the poorer quality ones, which are eliminated during the evolutionary process. Counting the frequency of use of each element at the primitive set can be considered as an impact index to reflect the relative importance of each input parameter. Hence, we are able to find what element is truly important and how important it could be for extracting C from satellite imagery.

Finally, we perform a linear regression analysis in order to map the GPVI into the C factor range. Hence, the C-values derived from the indices were compared to the C-values from the field measurements in order to validate the results. In addition, we perform a quantitative comparison between the RUSLE soil erosion estimation using the different sources of C as input. Except for the C factor, all other factors (R, K, L, S, and P) were determined using GIS software, according to the procedure described by Smith *et al.* (2007).

Experimental Results and Discussion

After a few exploratory runs required to fine tune the process configuration, 30 runs were performed to achieve the final configuration. The best individual produced at the end of each run was labeled as GPVI*j* (where *j* stands for the run number) and is considered as a candidate solution. Table 5 summarizes the performance of the solutions as measured by the difference between r_{train} and r_{test}. It can be seen that the GP approach is consistently able to identify reasonably VI's that generalizes well into the testing dataset. In general, the indices found by the GP process show better performance than the conventional indices (see Table 1 and Table 5). For example, the mean $\lvert r_{x,y} \rvert$ for GPVIs on the training dataset was 0.633 ± 0.01; whereas the mean $\lvert r_{x,y} \rvert$ for top ten conventional VI's on training dataset was 0.311 ± 0.01. Moreover all thirty GPVIs had a better $\lvert r_{x,y} \rvert$ than the best conventional index.

The structural analysis of the GPVIs (i.e., the components that made up the index) provides another interesting insight into the strategies identified by the evolutionary process to combine the primitive set. In this work we use a frequency of use (FOU) unit, which represents how often an element is used to generate a new index. Table 6 shows the frequency

TABLE 5. VI'S SYNTHESIZED BY THE GP PROCESS AND THEIR CORRELATION COEFFICIENT ($r_{x,\,y}$)

Index	Description	$\lvert r_{train} \rvert$	$\lvert r_{test} \rvert$	Diff.
GPVI1	$\rho_{\mathrm{green}} - (\rho_{\mathrm{SWIR1}} - \beta_{\mathrm{NIR}}) * \mathrm{I}_{\mathrm{RVI4}}$	0.620	0.730	0.111
GPVI2	$\beta_{\mathrm{SWIR1}}/(\mathrm{I}_{\mathrm{RVI4}}/(\mathrm{I}_{\mathrm{GEMI}}/\mathrm{I}_{\mathrm{RVI4}}))$	0.635	0.722	0.087
GPVI3	$\mathrm{NDSI}(\mathrm{I}_{\mathrm{RVI4}},\mathrm{I}_{\mathrm{GEMI}}) * (\mathrm{I}_{\mathrm{RVI4}} * \beta_{\mathrm{NIR}})$	0.619	0.681	0.062
GPVI4	$\mathrm{NDSI}(\mathrm{I}_{\mathrm{RVI4}}, \mathrm{NDSI}(\rho_{\mathrm{blue}},\mathrm{I}_{\mathrm{RVI4}}) * \mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}},\mathrm{I}_{\mathrm{RVI4}}))$	0.650	0.735	0.085
GPVI5	$(\mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}},\mathrm{I}_{\mathrm{RVI4}})/(\mathrm{I}_{\mathrm{RVI4}}/\beta_{\mathrm{SWIR1}})) * (\mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}},\mathrm{I}_{\mathrm{RVI4}})/\mathrm{I}_{\mathrm{RVI4}})$	0.638	0.743	0.105
GPVI6	$\mathrm{NDSI}(\mathrm{I}_{\mathrm{RVI4}},\mathrm{I}_{\mathrm{EVI}}) * (\mathrm{RSI}(\mathrm{I}_{\mathrm{RVI4}},\beta_{\mathrm{red}}) + \mathrm{I}_{\mathrm{RVI4}})$	0.615	0.682	0.067
GPVI7	$\mathrm{RSI}(\mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}},\mathrm{I}_{\mathrm{RVI4}}), (\rho_{\mathrm{red}} + 2\mathrm{I}_{\mathrm{RVI4}}))$	0.638	0.753	0.115
GPVI8	$\mathrm{NDSI}(\mathrm{NDSI}(\rho_{\mathrm{NIR}},\beta_{\mathrm{green}}) - \mathrm{I}_{\mathrm{RVI4}}, (\mathrm{I}_{\mathrm{GEMI}} - \rho_{\mathrm{blue}}) - \mathrm{I}_{\mathrm{RVI4}})$	0.643	0.673	0.030
GPVI9	$\mathrm{NDSI}(\mathrm{NDSI}(\mathrm{I}_{\mathrm{RVI4}},\beta_{\mathrm{green}}), \mathrm{I}_{\mathrm{RVI4}}) * \mathrm{I}_{\mathrm{GEMI}}$	0.620	0.724	0.104
GPVI10	$\mathrm{RSI}((\mathrm{I}_{\mathrm{GEMI}}/\mathrm{I}_{\mathrm{RVI4}})/(\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{EVI}}), b)$	0.635	0.707	0.073
GPVI11	$(\mathrm{I}_{\mathrm{GEMI}}/\mathrm{I}_{\mathrm{RVI4}})/\mathrm{I}_{\mathrm{RVI4}}$	0.633	0.773	0.140
GPVI12	$(\mathrm{I}_{\mathrm{RVI4}} - (\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{GEMI}})) - (\mathrm{I}_{\mathrm{GEMI}}/(\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{GEMI}}))$	0.638	0.752	0.115
GPVI13	$(\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{GEMI}}) - (\mathrm{I}_{\mathrm{GEMI}} - \rho_{\mathrm{green}})$	0.626	0.674	0.048
GPVI14	$\mathrm{RSI}(\mathrm{RSI}(\mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}},\mathrm{I}_{\mathrm{RVI4}}),(\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{EVI}})), (\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{EVI}}))$	0.638	0.717	0.079
GPVI15	$((\rho_{\mathrm{SWIR2}}/\mathrm{I}_{\mathrm{GEMI}}) + \rho_{\mathrm{green}}) + (\beta_{\mathrm{red}}/\mathrm{NDSI}(\mathrm{I}_{\mathrm{GEMI}},\mathrm{I}_{\mathrm{RVI4}}))$	0.645	0.733	0.088
GPVI16	$(\mathrm{I}_{\mathrm{GEMI}} - \mathrm{I}_{\mathrm{RVI4}}) - (\rho_{\mathrm{NIR}} - \mathrm{I}_{\mathrm{GEMI}})$	0.626	0.679	0.053
GPVI17	$\mathrm{RSI}(\rho_{\mathrm{NIR}},\mathrm{RSI}(\beta_{\mathrm{red}},\beta_{\mathrm{green}})) + \mathrm{RSI}(\mathrm{RSI}(\mathrm{I}_{\mathrm{RVI4}},\mathrm{a}),(\mathrm{I}_{\mathrm{GEMI}} - \mathrm{I}_{\mathrm{RVI4}}))$	0.648	0.645	0.002
GPVI18	$\mathrm{NDSI}((\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{GEMI}}),\mathrm{a}) * \mathrm{I}_{\mathrm{GEMI}}$	0.637	0.730	0.094
GPVI19	$((\mathrm{I}_{\mathrm{GEMI}} * \beta_{\mathrm{SWIR1}}) - \mathrm{I}_{\mathrm{GEMI}}) - \mathrm{I}_{\mathrm{RVI4}}$	0.622	0.629	0.008
GPVI20	$\mathrm{NDSI}(\rho_{\mathrm{NIR}},(\mathrm{I}_{\mathrm{RVI4}} + \mathrm{NDSI}(\rho_{\mathrm{blue}},\mathrm{I}_{\mathrm{GEMI}})))$	0.637	0.777	0.140
GPVI21	$\mathrm{NDSI}(\beta_{\mathrm{SWIR1}},\mathrm{I}_{\mathrm{GEMI}})/\mathrm{NDSI}(\mathrm{I}_{\mathrm{GEMI}},\mathrm{I}_{\mathrm{RVI4}})$	0.640	0.698	0.059
GPVI22	$(\mathrm{I}_{\mathrm{GEMI}} - b) - (\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{GEMI}})$	0.623	0.719	0.096
GPVI23	$\mathrm{RSI}((\rho_{\mathrm{blue}} - \mathrm{I}_{\mathrm{GEMI}}), \mathrm{I}_{\mathrm{RVI4}} * \mathrm{I}_{\mathrm{RVI4}} * b)$	0.648	0.710	0.062
GPVI24	$\mathrm{NDSI}((\mathrm{I}_{\mathrm{RVI4}} * \beta_{\mathrm{NIR}}),\rho_{\mathrm{NIR}}) * (\mathrm{I}_{\mathrm{RVI4}} * \beta_{\mathrm{NIR}})$	0.614	0.688	0.074
GPVI25	$\mathrm{I}_{\mathrm{RVI4}} * \beta_{\mathrm{SWIR1}} * (\beta_{\mathrm{NIR}} + \rho_{\mathrm{blue}})$	0.618	0.744	0.127
GPVI26	$\mathrm{RSI}(\mathrm{RSI}(\mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}}, \mathrm{I}_{\mathrm{RVI4}}), \mathrm{I}_{\mathrm{RVI4}}), \beta_{\mathrm{NIR}} * \mathrm{I}_{\mathrm{RVI4}})$	0.637	0.735	0.099
GPVI27	$\mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}}, \mathrm{I}_{\mathrm{RVI4}} - \rho_{\mathrm{NIR}}) - \rho_{\mathrm{NIR}}$	0.642	0.639	0.003
GPVI28	$\mathrm{RSI}(\mathrm{NDSI}(\rho_{\mathrm{red}},\mathrm{I}_{\mathrm{RVI4}}), \mathrm{RSI}(\mathrm{I}_{\mathrm{RVI4}}, \mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}},\mathrm{I}_{\mathrm{RVI4}})))$	0.643	0.655	0.013
GPVI29	$\mathrm{RSI}(\mathrm{RSI}(\mathrm{I}_{\mathrm{GEMI}}, \mathrm{I}_{\mathrm{RVI4}}), \mathrm{I}_{\mathrm{RVI4}})$	0.633	0.773	0.140
GPVI30	$(\mathrm{I}_{\mathrm{GEMI}}/(\mathrm{I}_{\mathrm{RVI4}} - \mathrm{I}_{\mathrm{GEMI}})) - \mathrm{I}_{\mathrm{GEMI}}$	0.638	0.752	0.115

TABLE 6. FREQUENCY OF INPUT PRIMITIVES USED IN THE GP PROCESS

Terminals		Functions	
Input	Frequency	Input	Frequency
ρ_{blue}	16.67%	'+'	16.67%
ρ_{green}	10.00%	'–'	46.67%
ρ_{red}	6.67%	'*'	36.67%
ρ_{NIR}	20.00%	NDSI	36.67%
ρ_{SWIR1}	3.33%	RSI	60.00%
ρ_{SWIR2}	3.33%		
β_{green}	10.00%		
β_{red}	10.00%		
β_{NIR}	16.67%		
β_{SWIR1}	16.67%		
I_{RVI1}	0.00%		
I_{RVI2}	0.00%		
I_{RVI4}	100.00%		
I_{RVI5}	0.00%		
I_{GEMI}	86.67%		
I_{NDVI}	0.00%		
I_{EVI}	10.00%		
a	6.67%		
b	10.00%		

of use of each element at the primitive set. In the case of elements belonging to Terminal set, it can be seen that I_{RVI4} has the highest FOU since it was used in all the 30 indices (i.e., 100 percent of the time). I_{RVI4} is followed by I_{GEMI}, which was used 86.67 percent of the time. In most cases, these elements appear together in all indices. By contrast, the rest of the elements had FOU values less than or equal to 20 percent. There were few elements that were not used at all (I_{RVI1}I, I_{RVI2}, I_{RVI5}, I_{NDVI}). Note that NDVI was not used by the GP process. This means that the GP algorithm identified that NDVI is not suitable to estimate C accurately. This statement will be illustrated below. Moreover, note that all the spectral bands (ρ_{blue}, ρ_{green}, ρ_{red}, ρ_{NIR}, ρ_{SWIR1}, ρ_{SWIR2}) were briefly used by the GP process. However, they are implicitly used on the conventional indices belonging to Terminal set. For example, I_{RVI4} contains ρ_{SWIR1} and ρ_{SWIR2}, while I_{GEMI} contains ρ_{NIR} and ρ_{red}. This is interesting because the former bands have been widely used to enhance plant moisture and lignin/cellulose components (Khana *et al.*, 2007; Streck *et al.*, 2002), whereas the latter have been widely used to detect green vegetation (Crist and Cicone, 1984; Jordan, 1969; Tucker, 1979). Twenty-eight out of the 30 GPVIs use ρ_{SWIR1}, ρ_{SWIR2}, ρ_{NIR} and ρ_{red}. This could indicate that the GP is able to identify spectral bands that are useful components to estimate the RUSLE C factor.

In the case of elements at the Functions set, the importance of each operator was more equally-distributed. The minus operator was used 46.67 percent of the time, whereas the plus operator was the least used (16.67 percent). The NDSI operator was used 36.67 percent of time, indicating that the GP recognized the NDVI structure as a useful structure to devise an index. This is interesting, inasmuch as the NDVI structure has been widely used in many scientific and operational applications (see reviews by Inoue and Olioso (2006); Moran *et al.* (1997)). The RSI operator was the most used: 60 percent of the time. This finding would indicate that the GP algorithm recognized the simple yet powerful properties of using a ratio format to eliminate a large proportion of the noise caused by changing sun angles, topography, clouds or shadow, and atmospheric conditions (Matsuchita *et al.*, 2007).

To validate the proposed approach, we compared the relation between the field measurements of C with the best three GPVI-derived C estimates. We considered that the most suitable solutions are those that show a similar correlation coefficient during the training and testing stages; hence, the best regional map of C can be expected from them. Therefore, according to the difference between r_{train} and r_{test}, the best three GPVIs obtained were GPVI17, GPVI27, and GPVI19, respectively. We also decided to include the best conventional index RVI4, the widely-used NDVI and the traditional cover classification method to compare the performance between all these approaches. Hence, we obtain C values from VI's by using linear regression analysis. For the cover classification method, a land cover map of the study area was obtained from the *Comisión Nacional Forestal* (CONAFOR; available at: **http://www.conafor.gob.mx**), and it is displayed as Figure 4. We assigned C values to labels that correspond to specific vegetative cover according to Table 10 from the USLE protocol (Wischmeier and Smith, 1978).

Figure 5 shows that the GPVI-based C factor calculations are the best related with field samples (GPVI17: $R^2 = 0.419$, RMSE = 0.017; GPVI27: $R^2 = 0.412$, RMSE = 0.017; GPVI19: $R^2 = 0.386$, RMSE = 0.016). On the other hand C-RVI4 obtains a modest performance ($R^2 = 0.2622$, RMSE = 0.019); whereas NDVI and the cover classification method have a poor performance in estimating C ($R^2 = 0.099$, RMSE = 0.035; and $R^2 = 0.073$, RMSE = 0.066, respectively). According to Figure 5, regression line of GPVI17 is the closest to the 1:1 line, which represents the field measurements of C. All indices show a tendency to underestimate C. For example, GPVI17 briefly underestimate C from $\sim$0.02 onwards; while GPVI27, GPVI19, RVI4, and NDVI underestimate C from $\sim$0.04 onwards. This fact could indicate that all indices tend to be more sensitive to the bare soil noise at low vegetation covers. In addition, all indices, except GPVI17 tend to overestimate C from 0 to $\sim$0.04 with different magnitudes. This fact could mean that those indices are not able to completely identify the elements that comprise C, yielding to C calculations higher than field measurements. Finally, the cover classification method completely overestimates the C factor, due to the fact that the homogeneity of the C factor estimates do not adequately reflect the spatial variation of field measurements.

Table 7 shows the averaged soil erosion obtained from testing dataset field locations, using the RUSLE model with different sources of C. It can be seen that the best approximations to field measurements are those derived from the synthetic indices. For example, the GPVI19-derived C factor has the best approximation to the field survey. Nevertheless, GPVI17- and GPVI27-derived C factors also yield a good approximation. RVI4-derived C obtains a modest performance because its prediction is $\sim$50 percent higher than the field survey. On the other hand, the NDVI and the Classification cover methods have a prediction that is twice the field measurements.

To illustrate the previous validation, we applied each VI's formula to the multi-spectral image in order to obtain their C factor maps (see Figure 6). We used the same adjustment criteria as Smith *et al.* (2007) to consider non-vegetative covers on the new C factor maps. For instance, negative values for C were set to 0. Moreover, C values greater than or equal to 0.45 within agricultural areas were assumed to be tilled and were set to C = 1.0. C values within water bodies were set to 0. Finally, urban areas (assumed to be largely paved or otherwise covered) were assigned C = 0.02.

Figure 6 shows five maps of C factor that correspond to linear regressions of GPVI17, GPVI27, GPVI19, RVI4, NDVI, which have a range from 0.0 to 1.0 across the watershed, and a mean of 0.0275 ± 0.0252; 0.0224 ± 0.0192; 0.0297 ± 0.0243; 0.0321 ± 0.0280; and 0.0385 ± 0.0356,

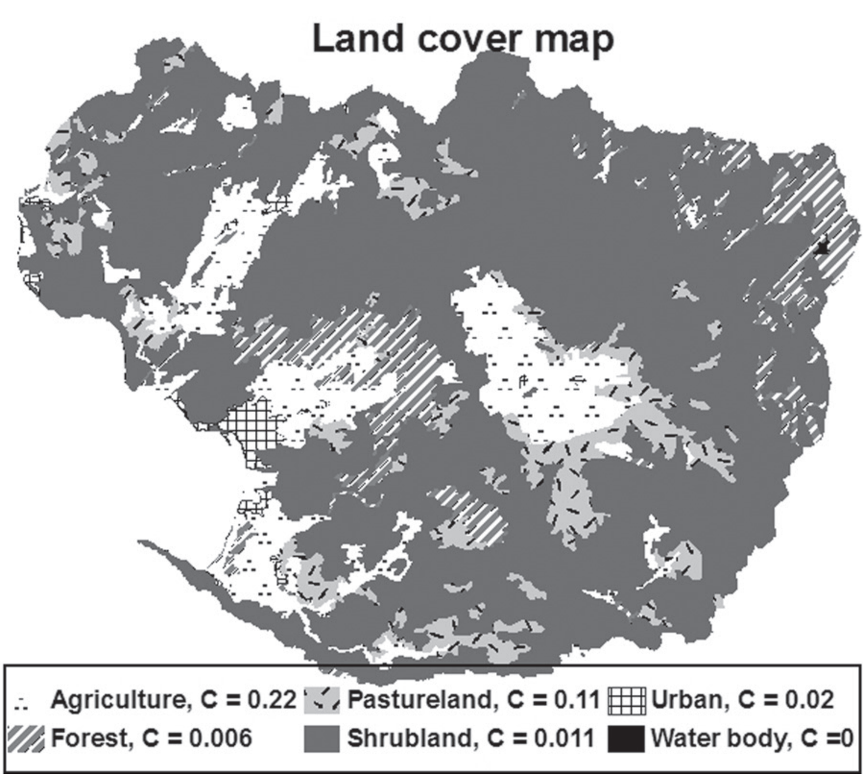


Figure 4. General land cover map for the watershed. This map includes the assigned C-factor values for the land cover types

respectively. C factor values were grouped into 10 classes for comparative purposes, whereby classes ranged from 0 (water and full vegetation) to 1 (bare soil). According to on-field and imagery inspections, the study area is mostly dominated by shrub-land, pasture land, and agriculture. However there are some urban settlements at the west, whereas some forest patches are depicted at the east.

At a general glance, all maps in Figure 6 seem to be satisfactory, in regard to the main land cover units of the study area. However, looking more closely at the maps, the number of pixels assigned to certain classes gives a more detailed picture. Note that a clear distinction over all maps can be seen for the C factor values ranging from 0.0 to 0.001, which represent highly-covered mountain forest. The relative distribution of the pixels over all maps shows two peaks: one at the first class and the other between fourth and fifth classes. These finding would indicate that a greater portion of the study area has C factor values ranging from 0 to 0.001 and 0.01 to 0.05. Note that at that range GPVIs 17, 19, and 27 are the closest indices to field measurements according to Figure 5 and Table 7. Hence, the C maps from GPVIs are more reliable than C maps from conventional indices. On the other hand, the NDVI-based map shows a briefly different tendency. It has the second peak between the fifth and the sixth class. This fact can be explained by the marked tendency of NDVI to overestimate C-values for green vegetation (from 0 to 0.04) and to underestimate C-values for the litter layer and vegetation under stress conditions (from 0.04 onwards). This phenomenon is revealed in Figure 6e. It can be seen that NDVI is not sensitive to the dry pasture land at the center of the watershed. By contrast, the rest of the maps enhance this area because they use information from the shortwave infrared bands, which are able to discriminate dry plant matter from bare soil (Khana *et al.*, 2007). However, using information from short wave bands only do not assure a good estimation of C factor, as can be seen on the RVI4-based map (Figure 6d). Below the dry pasture land at the center of the watershed, a sparse, dry shrub land exists that RVI4 cannot completely discriminate from bare soil. However, all GPVI-based maps can enhance this and others similar areas because they use more information from ρ_{NIR}, ρ_{red} besides ρ_{SWIR1}, $\rho_{\mathrm{SWIR2,}}$, reflective bands.

Conclusions and Future Work

In this work a novel genetic programming approach was described, which automatically creates vegetation indices that have a good correlation with the RUSLE C factor. In this way, several new indices were designed that result in improved correlation coefficients with C factor field

Table 7. Mean and Standard Deviation of the Soil Erosion Rates for Testing Dataset using Different C Estimations

Method used to estimate C	Soil erosion in Mg km^{-2} yr.$^{-1}$
Field survey	76.6 ± 157.4
GPVI17	67.1 ± 97
GPVI19	76.5 ± 103.5
GPVI27	68.6 ± 97.4
RVI4	119.7 ± 184.4
NDVI	193.7 ± 393.3
Cover Classification	170.6 ± 358.8

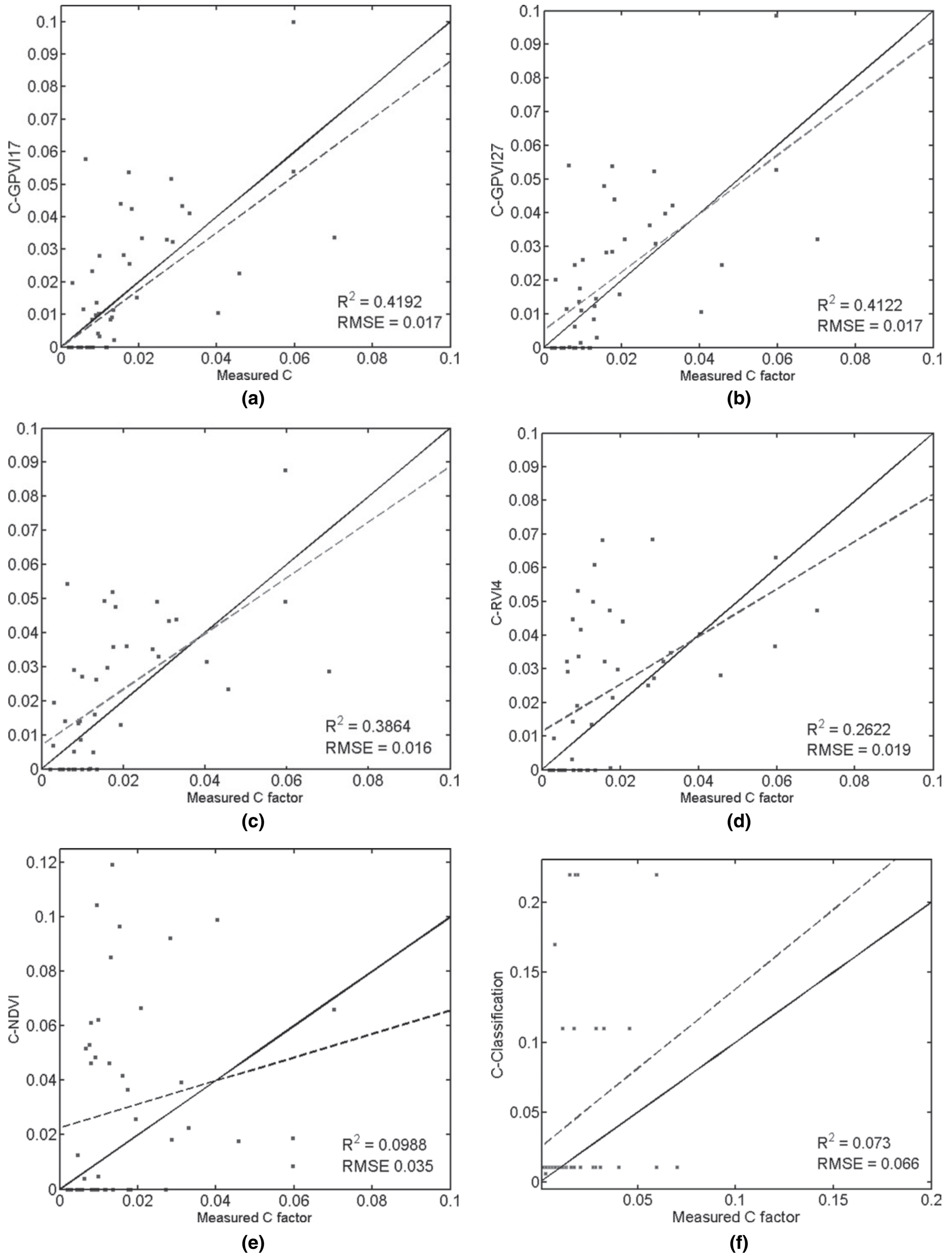
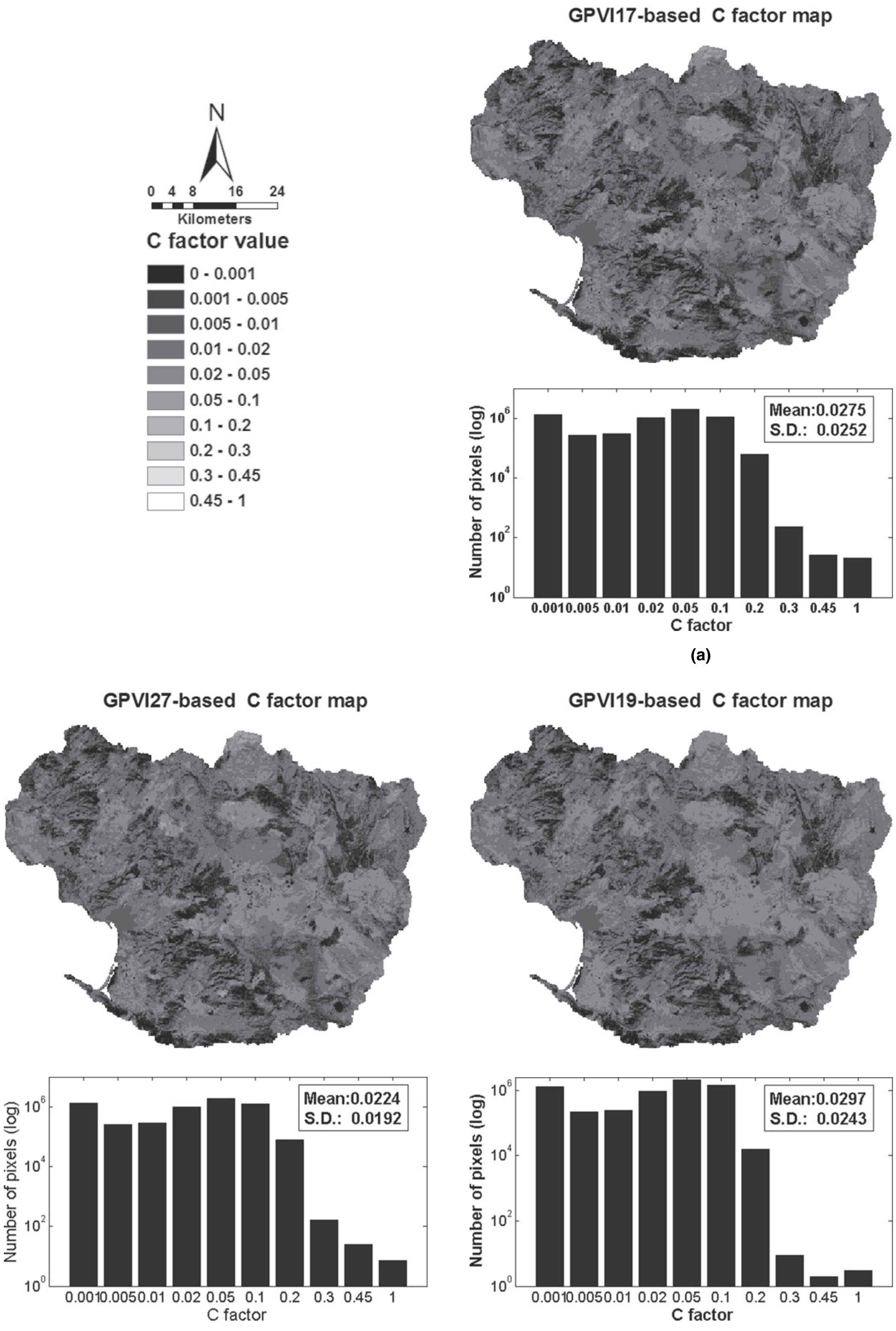


Figure 5. Comparison of the C factor measured in the field against calculated C factor using (a) GPVI17, (b) GPVI27, (c) GPVI19, (d) RVI4, (e) NDVI, and (f) Cover classification method. The solid line is the 1:1 line, and the dashed line is the regression line.

samples, and actually it gives reasonable better performance than widely-used indices such as NDVI and an alternative to RVI, the so-called RVI4.

Thus, we introduce several new indices called GPVI*j,* from which we distinguish as the best GPVI17, GPVI27, and GPVI19. Such indices achieved $|r_{x,y}|$ values during the training stage of 0.65, 0.64, and 0.62, respectively; while the conventional indices NDVI and RVI4 achieved a $|r_x,y|$ value for the training stage of 0.31 and 0.51, respectively. In general, all thirty GPVIs had a better correlation coefficient than the best conventional index. Hence, our results demonstrate that GP is a useful tool to find out different band combinations that are able to identify the basic elements to estimate the RUSLE C factor. In particular, our



(a)

(b)

(c)

Figure 6. C factor maps derived using (a) GPVI17, (b) GPVI27, and (c) GPVI19. The figure also shows the pixel distribution for each map.

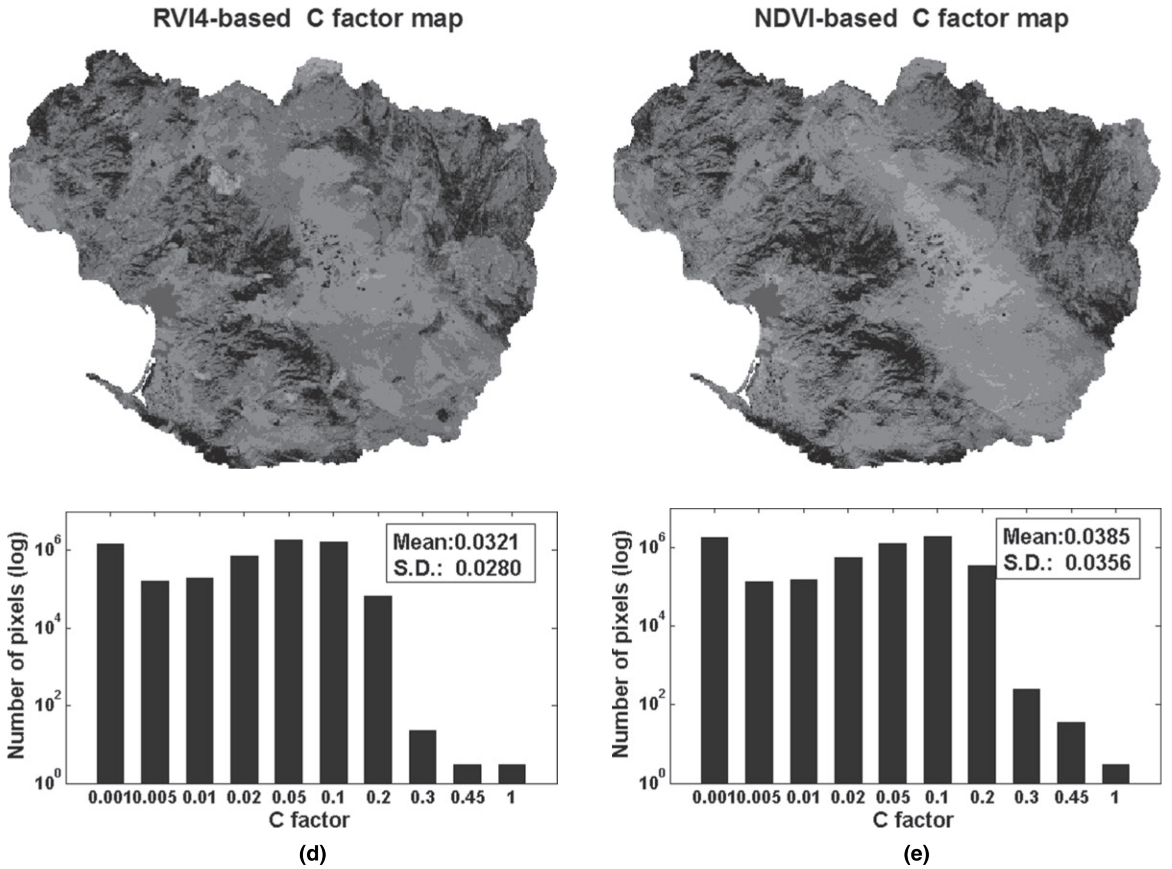


Figure 6 continued. C factor maps derived using (d) RVI4, and (e) NDVI. The figure also shows the pixel distribution for each map.

GP approach was able to identify ρ_{SWIR1}, ρ_{SWIR2}, ρ_{NIR} and ρ_{red} as the most useful bands for extract the RUSLE C factor from satellite imagery. Moreover, GP was able to recognize the simple yet powerful property of using the ratio (RSI) and the NDSI structures to eliminate a large proportion of the noise caused by changing sun angles, topography, clouds or shadow, and atmospheric conditions. Hence, this paper shows that our approach obtain fairly good results on estimating C for a particular watershed (Todos Santos). However, the generality of our approach depends on further application to other watersheds, which is part of our future work.

Based on our experimental results, we believe that the data mining capability of the GP makes this approach ideally suited for implementing multi-spectral analysis and other remote sensing applications, such as classifiers and mathematical indices that enhance the signal of several features over the Earth's surface. Moreover, the advantage of the GP approach is its open box characteristic. If a black box where used (e.g., neural networks, fuzzy logic, and most statistical approaches), we could hardly find out what input parameter is truly important and how important it could be.

Our future work has two main purposes. The first purpose is related to the generality of our approach to obtain a good index to estimate C. The generality of our approach depends on its further application to other watersheds. We believe that as we keep applying our approach to different field sites, we could identify a tendency on the elements that comprises the best indices; or even identify a single index that could perform well in general. This information would give us certainty on that index, and then would not require us to synthesize new indices. We encourage the use of the indices reported on this paper to estimate C, so we can obtain a feedback to generalize on different conditions, and we can find a single index that could be applied to estimate the C factor in general. The second purpose will focus on estimating the soil erosion rate in Todos Santos watershed with the RUSLE model; based on the C factor maps obtained from this work. Table 7 shows that the best approximations to field measurements are those derived from the synthetic indices. However, we understand that these results apply for the data samples only. We will have to carry out an extensive field campaign to strongly demonstrate that our improved estimations of C could yield to an improvement on the prediction of soil erosion rates. Nevertheless, the results presented here give us a useful insight into how the improvements on the C factor could be extended over the entire watershed; and encourage us to keep exploring this research avenue.

Acknowledgments

This research has been supported by CICESE internal funds. The first author would like to thank the Mexican Council for Science and Technology (CONACYT) for extending financial support in the conduct of his research. We are grateful to anonymous reviewers for their careful reviews that improve the manuscript.

References

Agnelli, D., A. Bollini, and L. Lombardi, 2002. Image classification: An evolutionary approach, *Pattern Recognition Letters*, 23:303–309.

Asis, A.M., and K. Omasa, 2007. Estimation of vegetation parameter for modeling soil erosion using linear spectral mixture analysis of Landsat ETM data, *ISPRS Journal of Photogrammetry and Remote Sensing*, 62:309–324.

Asner, G.P., and D.B. Lobell, 2000. A biogeophysical approach for automated SWIR unmixing of soils and vegetation, *Remote Sensing of Environment*, 74:99–112

Bauer, H.L., 1943. The statistical analysis of chaparral and other plant communities by means of transect samples, *Ecology*, 24:45–60.

Bhanu, B., and Y. Lin, 2003. Genetic algorithm based feature selection for target detection in SAR images, *Image and Vision Computing*, 21:591–608.

Brady, N.C., and R.R. Weil, 2008. *The Nature and Properties of Soils*, Fourth edition, Pearson Prentice Hall, New Jersey.

Chander, G., and B. Markham, 2003. Revised LANDSAT-5 TM radiometric calibration procedures and post-calibration dynamic ranges, *IEEE Transactions on Geoscience and Remote Sensing* 41(3):2674–2677.

Crist, E.P. and R.C. Cicone, 1984. Application of the tasseled cap concept to simulated Thematic Mapper data, *Photogrammetric Engineering & Remote Sensing*, 50(3):343–352.

Daida, J.M., J. Hommes, T. Bersano-Begey, S. Ross, and J. Vesecky, 1996. *Algorithm Discovery using the Genetic Programming Paradigm: Extracting Low-contrast Curvilinear Features from SAR Images of Arctic Ice*, MIT Press, Cambridge, Massachusetts.

Folly, A., M.C. Bronsveld, and M. Clavaux, 1996. A knowledge-based approach for C-factor mapping in Spain using Landsat TM and GIS, *International Journal of Remote Sensing*, 17(2):2401–2415.

Gao, B., 1996. NDWI - A normalized difference water index for remote sensing of vegetation liquid water from space, *Remote Sensing of Environment*, 58:257–266.

Guerschman, J.P., M.J. Hill, L.J. Renzullo, D.J. Barrett, A.S. Marks, and E.J. Botha, 2009. Estimating fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil in the Australian tropical savanna region upscaling the EO-1 Hyperion and MODIS sensors, *Remote Sensing of Environment*, 113:928–945.

Harvey, N., J. Thelier, S. Brumby, S. Perkins, J. Szymanski, J. Bloch, R. Porter, M. Galassi, and A. Young, 2002. Comparison of GENIE and conventional supervised classifiers for multispectral image feature extraction, *IEEE Transactions on Geoscience and Remote Sensing* 40(2):393–404.

Howard, D., S.C. Roberts, and R. Brankin, 1999. Target detection in SAR Imagery by Genetic Programming, *Advances in Engineering Software*, 30:303–311.

Huete, A.R., 1988. A Soil-Adjusted Vegetation Index (SAVI), *Remote Sensing of Environment*, 25:295–309.

Inoue, Y., J. Peñuelas, A. Miyata, and M. Mano, 2008. Normalized difference spectral indices for estimating photosynthetic efficiency and capacity at a canopy scale derived from hyperspectral and CO_2 flux measurements in rice, *Remote Sensing of Environment*, 112(1):156–172.

Inoue, Y., and A. Olioso, 2006. Estimating dynamics of ecosystem CO2 flux and biomass production in agricultural field by synergy of process model and remotely sensed signature, *Journal of Geophysical Research D, Atmospheres*, 111:D24S91, doi:10.1029/2006JD007469.

de Jong, S.M., 1994. *Applications of Reflective Remote Sensing for Land Degradation Studies in a Mediterranean Environment*, Ph.D. dissertation, Universiteit Utrecht, Nederlands.

Jordan, C.F., 1969. Derivation of leaf area index from quality of light on the forest floor, *Ecology*, 50:663–666.

Khana, S., A. Palacios-Orueta, M.L. Whiting, S.L. Ustin, D. Riaño, and J. Litago, 2007. Development of angle indexes for soil moisture estimation, dry matter detection and land-cover discrimination, *Remote Sensing of Environment*, 109:154–165.

Koza, J.R., 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press Cambridge, Massachusetts.

Koza, J.R., 2010. Human-competitive results produced by genetic programming, *Genetic Programming and Evolvable Machines*, 11(3–4):251–284.

Lin, C.Y., W.T. Lin, and W.C. Chou, 2002. Soil erosion prediction and sediment yield estimation: The Taiwan experience, *Soil and Tillage Research*, 68(2):143–152.

Lu, H., I.P. Prosser, C.J. Moran, J.C. Gallant, G. Priestly, and J.G. Stevenson, 2003. Predicting sheetwash and rill erosion over Australian continent, *Australian Journal of Remote Sensing*, 41(6):1037–1062.

Makkeasorn, A., N. Chang, and J. Li, 2009. Seasonal change detection of riparian zones with remote sensing images and genetic programming in a semi-arid watershed, *Journal of Environment Management*, 90:1069–1080.

Matsuchita, B., W. Yang, J. Chen, Y. Onda, and G. Qiu, 2007. Sensitivity of the Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) to topographic effects: A case study in high-density cypress forest, *Sensors*, 7:2636–2651.

Moran, M.S., Y. Inoue, and E.M. Barnes, 1997. Opportunities and limitations for image-based remote sensing in precision crop management, *Remote Sensing of Environment*, 61:319–346.

Olague, G., 2002. Automated photogrammetric network design using genetic algorithms, *Photogrammetric Engineering & Remote Sensing*, 68(5):423–431.

Pinty, B., and M.M. Verstraete, 1992. GEMI: A non-linear index to monitor global vegetation from Satellites, *Vegetation*, 101:15–20.

Poli, R., W. Langdon, and N. Freitag, 2008. A field guide to genetic programming, URL: *http://www.gp-field-guide.org.uk* (last date accessed: 21 January 2011).

Rauss, P., J. Daida, and S. Chaudhary, 2000 Classification of spectral imagery using genetic programming, *Proceedings of Genetic and Evolutionary Computation Conference*, 08–12 July, Las Vegas, Nevada, pp. 726–733.

Renard, K.G., G.R. Foster, G.A. Weesies, D.K. McCool, and D.C. Yoder, 1997. *Predicting Soil Erosion by Water: A Guide to Conservation Planning with the Revised Universal Soil Loss Equation*, Agricultural Handbook, Number 703, U.S. Government Printing Office, Washington, D.C.

Ross, B.J., A.G. Gualtieri, F. Fueten, and P. Budkewitsch, 2005. Hyperspectral image analysis using genetic programming, *Applied Soft Computing*, 5:147–156.

Rouse, J.W., R.H. Haas, J.A. Schell, and D.W. Deering, 1973. Monitoring vegetation systems in the great plains with ERTS, *Proceedings of the Third ERTS Symposium*, NASA SP-351, 1:309–317.

Ryerson, R.A. (editor), 1999. *Manual of Remote Sensing, Third edition, Volume 3: Remote Sensing for the Earth Sciences*, American Society for Photogrammetry and Remote Sensing, John Wiley & Sons.

Sabins, F.F., 1997. *Remote Sensing: Principles and Interpretation, Third edition*, Freeman, New York, 496 p.

Silva, S., 2007. GPLAB - A GP Toolbox for MATLAB, URL: *http://gplab.sourceforge.net/index.html* (last date accessed: 21 January 2011).

Smith, S. V., S.H. Bullock, A. Hinojosa-Corona, E. Franco-Viscaíno, M. Escoto-Rodríguez, T.G. Kretzschmar, L.M. Farfan, and J.M. Salazar-Cesena, 2007. Soil erosion and significance for carbon fluxes in a mountainous Mediterranean-climate watershed, *Ecological Applications*, 17(5):1379–1387.

Streck, N.A., D. Rundquist, and J. Connot, 2002. Estimating residual wheat dry matter from remote sensing measurements, *Photogrammetric Engineering & Remote Sensing*, 68(11):1193–1201.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation, *Remote Sensing of Environment*, 8:127–150.

Wang, G., S. Wente, G. Gertner, and A. Anderson, 2002. Improvement in mapping vegetation cover factor for the universal soil loss equation by geostatistical methods with Landsat Thematic Mapper images, *International Journal of Remote Sensing*, 23(18):3649–3667.

Weltz, M.A., K.G. Renard, and J.R. Simanton, 1987. Revised universal soil loss equation for western rangeland, *Proceedings of the Symposium of Strategies for Classification and Management of Native Vegetation for Food Production in Arid Zones,* USDA-GTR, RM-150, Tucson, Arizona, pp. 104–111.

Wischmeier, W.H., and D.D. Smith, 1978. *Predicting Rainfall Erosion Losses: A Guide to Conservation Planning*, Agricultural Handbook, Number 537, U.S. Government Printing Office, Washington, D.C.

Vincent, R.K., 1997. *Fundamentals of Geological and Environmental Remote Sensing*, Prentice Hall, Upper Saddle River, New Jersey, 370 p.

Zippin, D.B., and J.M. Vanderwier, 1994. Scrub community descriptions of the Baja California Peninsula, Mexico, *Madroño*, 41:85–119.

(Received 27 January 2010; accepted 23 April 2010; final version: 28 November 2010)