Deriving Vegetation Indices for Phenology Analysis using Genetic Programming

Jurandy Almeida$^{a,d,*}$, Jefersson A. dos Santos$^b$, Waner O. Miranda$^b$, Bruna C. Alberton$^c$, Leonor Patricia C. Morellato$^c$, Ricardo da S. Torres$^d$

$^a$ Institute of Science and Technology, Federal University of São Paulo – UNIFESP
12247-014, São José dos Campos, SP – Brazil

$^b$ Department of Computer Science, Universidade Federal de Minas Gerais – UFMG
31270-010, Belo Horizonte, MG – Brazil

$^c$ Phenology Lab, Dept. of Botany, São Paulo State University – UNESP
13506-900, Rio Claro, SP – Brazil

$^d$ RECOD Lab, Institute of Computing, University of Campinas – UNICAMP
13083-852, Campinas, SP – Brazil

Abstract

Plant phenology studies recurrent plant life cycle events and is a key component for understanding the impact of climate change. To increase accuracy of observations, new technologies have been applied for phenological observation, and one of the most successful strategies relies on the use of digital cameras, which are used as multi-channel imaging sensors to estimate color changes that are related to phenological events. We monitor leaf-changing patterns of a cerrado-savanna vegetation by taking daily digital images. We extract individual plant color information and correlate with leaf phenological changes. For that, several vegetation indices associated with plant species are exploited for both pattern analysis and knowledge extraction. In this paper, we present a novel approach for deriving appropriate vegetation indices from vegetation digital images. The proposed method is based on learning phenological patterns from plant species through a genetic programming framework. A comparative analysis of different vegetation indices is conducted and discussed. Experimental results show that our approach presents higher accuracy on characterizing plant species phenology.

Keywords: remote phenology, digital cameras, image analysis, vegetation indices, genetic programming

1. Introduction

Phenology is the study of recurrent natural phenomena and its relationship to climate (Schwartz, 2013). Traditional phenology studies rely on the direct observation of plants, which is a time consuming and error-prone task. Recently, digital cameras have been applied as tools to monitor leaf changes on plants automatically (Alberton et al., 2014; Morellato et al., 2013; Schwartz, 2013).

The digital cameras can increase the accuracy of phenological observation and widen the study area, but its successful application as multi-channel imaging sensors to capture vegetation changes is reliant to the extraction of color change information out of the images (Alberton et al., 2014; Sonnentag et al., 2012). Generally, leaf color information is extracted from the red, blue, and green (RGB) color channels and the green channel is the most utilized to describe leaf changes, in combination to red color (Richardson et al., 2009, 2007). The normalized RGB chromatic coordinates are currently considered the most reliable indices to describe phenological changes based on image analyses (Sonnentag et al., 2012).

Considering the actual relevance of phenology as a tool for monitoring plant responses to climatic change and the need to reveal the environmental triggers of tropical phenology, here we propose a novel approach for deriving appropriate vegetation indices from vegetation digital images. The proposed method is based on learning phenological patterns from plant species through a Genetic Programming (GP) framework (da S. Torres et al., 2009). According to this framework, complex combinations of vegetation indices are modeled as individuals of a population. These individuals are then evolved by means of genetic operators (e.g., crossover, mutation, and reproduction) along generations. The objective is to obtain better performing complex vegetation indices that can be used to characterize the behavior of plant species or functional groups.

We performed a rigorous comparative analysis of different vegetation indices and discussed our proposed index accuracy in relation to the most utilized in the literature, the chromatic coordinates index, for near surface remote phenology studies (e.g., Ahrends et al. (2009); Alberton et al. (2014); Nagai et al. (2011); Richardson et al. (2009, 2007); Sonnentag et al.)
(2012)). Our experimental results show that vegetation indices may complement each other and improve indices’ accuracy on characterizing plant species.

The remainder of this paper is organized as follows. Section 2 describes the methodology adopted for acquiring time series. Section 3 briefly discusses vegetation indices utilized in phenology analysis. Section 4 presents the GP framework and shows how to apply it to model complex vegetation indices. Section 5 describes the experimental protocol adopted for evaluating vegetation indices. Section 6 reports the results of our experiments and compares our proposed index with other ones. Finally, we offer our conclusions and directions for future work in Section 7.

2. Time Series Acquisition

A digital hemispherical lens camera (Mobotix Q24) was set up in a 18m tower in a Cerrado sensu stricto, a savanna vegetation located at Itirapina, São Paulo State, Brazil. We set up the camera to take a daily sequence of five JPEG images (at 1280 × 960 pixels of resolution) per hour, from 6:00 to 18:00 h (UTC-3). The present study was based on the analysis of over 2,700 images, recorded at the end of the dry season, between August 29th and October 3rd 2011, day of year 241 to 278, during the main leaf flushing season (Alberton et al. 2014, Almeida et al. 2014).

The image analysis was conducted by defining different regions of interest (ROI), as described in Richardson et al. (2009, 2007) and defined by Alberton et al. (2014) for our target species (Figure 1). We analyzed 22 ROIs (Figure 1) of six plant species randomly selected in the hemispheric image: (i) three regions associated with Aspidosperma tomentosum (red areas), (ii) four regions for Caryocar brasiliensis (green areas), (iii) two regions for Myrica guianensis (blue areas), (iv) seven regions for Miconia rubiginosa (orange areas), (v) two regions for Pouteria ramiflora (magenta areas), and (vi) four regions for Pouteria torta (cyan areas)

According to the leaf exchange data from the on-the-ground field observations on leaf fall and leaf flush at our study site, those species were classified into three functional groups (Albert et al. 2014, Morellato et al. 1989): (i) deciduous, A. tomentosum and C. brasiliensis; (ii) evergreen, M. guianensis and M. rubiginosa; and (iii) semideciduous, P. ramiflora and P. torta.

3. Vegetation Index-based Phenology Analysis

Digital images allow the detection of phenological events according to the changes of red, green, and blue (RGB) color channels along time (Richardson et al. 2007). By quantifying the RGB color channels it is possible to estimate, for instance, leaf flushing and senescence, using the green and red channels, respectively (Ahrends et al. 2009, Henneken et al. 2013, Morisette et al. 2009, Richardson et al. 2009). The quantification of RGB is performed applying indices of color channels to detect the leaf color changes in temporal time series of digital images (Nagai et al. 2011, Richardson et al. 2007, Sonnentag et al. 2012, Zhao et al. 2012, Zhou et al. 2013).

Non-normalized RGB coordinates correspond to the individual red (R), green (G), and blue (B) components of a digital image of 24 bit color (8 bits for each color). These values can be highly sensitive to the intensity and variation of the illuminating source and its angle with the background (Gonzalez and Woods 2007). Non-normalized RGB coordinates tend to present considerable variation between images, specially related to light variation in shaded and not shaded surfaces (Woebbecke et al. 1995, Zhou et al. 2013) tested non-normalized values for the RGB color channels (RGBDN) and the indices fail to capture any seasonal change in the development of the canopy of winter wheat cultivated area.

The contrast indices can detect optical contrast between plants and non-plant surfaces (background) in an image (Woebbecke et al. 1995). The Excess green (2G − R − B) is a commonly applied contrast index to highlight the green information and complement phenological interpretation in several studies (e.g., Kurc and Benton 2010, Migliavacca et al. 2011, Nagai et al. 2011, Sonnentag et al. 2012). The G/R, calculated on the basis of the difference of absorptive/reflective bands, corresponding to the vegetation canopy and soil surface, provides effective vegetative information for leaf color changes in Zhou et al. (2013). Other examples of contrast indices are R − G, G − B, and (G − B)/(R − G) (Woebbecke et al. 1995).

The perceptive attributes of color are classified as hue, saturation, and intensity (HSI model) (Woebbecke et al. 1995). Hue (or spectral shape) is measured as the wavelength of maximal reflection (λRmax). Hue modeling describes “redness”, “greenness”, and “blueness” of an object and it is calculated by the non-normalized RGB color channels (Woebbecke et al. 1995). Methods calculating Hue index were useful to detect leaf-colour and leaf-fall patterns and their timings among...
species in deciduous broad-leaved trees (Nagai et al., 2011).

The normalized index called RGB chromatic coordinates (RGBcc) was developed by Gillespie et al. (1987) and it is considered the most efficient index to detect the color of plants in relation to their background (Woebbecke et al., 1995). Shaded and unshaded surfaces presented less variation in relation to other non-normalized indices (Woebbecke et al., 1995). Normalized chromatic coordinates is the most efficient index to distinguish leaves between monocots and dicots and to suppress light environment variation (see Alberton et al., 2014; Sonnentag et al., 2012; Woebbecke et al., 1995). The normalized RGBcc is calculated according to Gillespie et al. (1987) and Woebbecke et al., (1995). as follow:

\[ r_{cc} = \frac{R}{(R + G + B)} \] (1)
\[ g_{cc} = \frac{G}{(R + G + B)} \] (2)
\[ b_{cc} = \frac{B}{(R + G + B)} \] (3)

4. Vegetation Index Discovery Process

This section introduces the proposed vegetation index discovery process. Section 4.1 provides a background on GP concepts, while Section 4.2 presents how GP can be used to create complex vegetation indices.

4.1. The GP Framework

Genetic programming (GP) (Koza, 1992) is a set of artificial intelligence search algorithms designed according to biological evolution principles. A solution for a target problem is modeled as an individual of a population. Population of individuals are then evolved along generations with the objective of creating better individuals (i.e., better solutions) in subsequent generations. Different genetic operations, such as reproduction, crossover, and mutation, are employed to improve a population. Trees are the most common representation of individuals in a GP framework. Leaf nodes are associated with data and inner nodes are related to operations. Figure 2 shows an example of an individual. In this figure, leaf nodes are variables \( x, y, \) and \( z \); and inner nodes are mathematical operators +, *, /, and sqrt, resulting in the function \(((x/y) + x)*/(sqrt(z)))\).

Usually, a random approach is employed for creating an initial set of individuals using tree representations (Koza, 1992). Once an initial population is generated, the evolutionary process is started. The first step consists in assessing the quality of an individual, i.e., a solution. For that, a fitness function is used to assign a fitness value for each individual. Those individuals with better performing scores are then selected to take part in the population of the next generation. Furthermore, other populations are created by applying genetic transformation operations, such as reproduction, crossover, and mutation. The reproduction operator identifies the best individuals and copies them to the next generation. The crossover operator, in turn, swaps subtrees of two parents, i.e., their genetic material. The mutation operator selects a node of a tree (usually randomly) and replaces its subtree by a new randomly generated subtree. More details regarding the implementation of these operators can be found in (Koza, 1992; Muttil and Lee, 2005).

Algorithm 1 presents the overall framework. Starting with a set of training data with known relevance judgements, GP first creates a population of random combination functions (Line 1). These combination functions are then evaluated based on the relevance information available in the training data. If a predefined stopping criteria is not reached, the genetic transformations are applied with the objective of creating the next population generation iteratively (Lines 2–8). Finally, after a number of generations, the individual with the highest fitness value, i.e., the solution that best solves the problem, is chosen and applied on a set of testing data (Line 9).

Algorithm 1 GP Framework

1: Generate an initial population of random “individuals”
2: while number of generations \( \leq N_{gen} \) do
3: Calculate the fitness of individual
4: Record the top \( N_{top} \) individuals
5: for all the \( N_{top} \) individuals do
6: Create the new population by the reproduction, crossover and mutation operations.
7: end for
8: end while
9: Apply the “best individual” (i.e., the first tree of the last generation) on a set of testing (query) data

In this work, GP is used to find the best possible combination from the available vegetation indices. We chose GP for a number of reasons, including: (i) excellent effectiveness in previous document retrieval studies (Andrade et al., 2012; Calumby et al., 2014; da S. Torres et al., 2009; Faria et al., 2010; Ferreira et al., 2011), mainly when exploiting multiple information; (ii) capability to find near-optimal solutions in large search spaces, as is the case here; (iii) and capability to provide easy-to-interpret complex combination functions that are associated with possibly unexpected novel vegetation indices. In fact, as far as we know, GP has never been applied in the scenario we deal with in this article. With regard to other learning approaches such as neural networks, Muttil and Lee have raised the following advantages of using GP (Muttil and Lee, 2005):

![Figure 2: Example of a tree representation of an individual in a GP framework.](image-url)
GP does not assume any a priori functional form of the solution; GP models may provide additional insights into the problem at hand as they provide analytical equations that can be interpreted; and GP can reduce substantially the dimensionality of the model, with consequent benefits on model interpretation.

### 4.2. Modeling Complex Vegetation Indices as GP Individuals

In this paper, we adopted the GP framework proposed by da S. Torres et al. (2009) for combining image descriptors to support queries based on visual content. A descriptor $D$ is modeled as a tuple $(\epsilon, \delta)$, where $\epsilon$ is a feature extraction function and $\delta$ is a distance function. In our problem, $\epsilon$ refers to a function that extracts vegetation index time series for image regions.

For example, let $D_{\epsilon_r} = (\epsilon_{r,}, \delta_{r,})$, $D_{\epsilon_g} = (\epsilon_{g,}, \delta_{g,})$, and $D_{\epsilon_b} = (\epsilon_{b,}, \delta_{b,})$ be three descriptors whose feature extraction functions $\epsilon_{r,}, \epsilon_{g,}$, and $\epsilon_{b,}$ extract, from a given vegetation image region, the time series related to the normalized contribution over time of the Red, Green, and Blue channels, respectively, according to Equations 1, 2, and 3 presented in Section 3. Each time series is seen as a feature vector that characterizes an image region, as illustrated in Figure 3.

The resulting time series are concatenated into a single vector representation. In the following, a new similarity function can now be used to compute the similarity between two image regions $\hat{A}$ and $\hat{B}$, by using their time series (feature vector) representations.

Considering our previous example, assume that the similarity functions $\delta_{r,}$, $\delta_{g,}$, and $\delta_{b,}$ are implemented as the inverse of the Manhattan distance (say $\delta$) of the times series computed from the two regions $\hat{I}_A$ and $\hat{I}_B$, i.e., $\delta = \delta_{r,}(\hat{I}_A, \hat{I}_B) = \delta_{g,}(\hat{I}_A, \hat{I}_B) = \delta_{b,}(\hat{I}_A, \hat{I}_B)$. Figure 4 illustrates the use of the GP individual depicted in Figure 2 as a similarity function that combines three different vegetation index descriptors ($D_{\epsilon_r} = (\epsilon_{r,}, \delta_{r,})$, $D_{\epsilon_g} = (\epsilon_{g,}, \delta_{g,})$, and $D_{\epsilon_b} = (\epsilon_{b,}, \delta_{b,})$).

Usually, the computation of the GP-based vegetation index is performed completely offline and, hence, does not impact on the search time.

### 5. Experimental Protocol

Unlike other research areas, evaluating a vegetation index in the context of phenoology is not a straightforward task due to the lack of an objective ground-truth. Nowadays, some indices have been tested in the framework of near-surface remote pheno-logy studies (e.g., Sonnentag et al., 2012).

In this work, we adopted the evaluation method used in Almeida et al. (2013a,b). It relies on the identification of plant species in the image using time series extracted from pixels associated with individuals of a same species. This evaluation method addresses the plant identification as an image retrieval problem, in contrast to some initiatives in which this task is addressed in the context of image classification (Almeida et al., 2012, 2014). The flowchart of the image retrieval system is illustrated in Figure 5. In this system, we provide a time series extracted from an image area associated with a given species and we query for similar time series computed from other image areas that belong to the same species.

For that, the Guigues algorithm (Guigues et al., 2006) was used to segment the hemispheric image into small polygons, obtaining 8,849 segmented regions (SR). Then, each SR was
associated with a single ROI aiming to label it. A labeled region is created if there is at least 80% of overlapped area between a SR and a ROI. In the remainder of this paper, when we refer to segmented regions obtained by the Guigues algorithm, we use the acronym SR; whereas when we refer to regions of interest related to tree crowns of plant species identified manually in the digital image, we use the acronym ROI. In this way, we built a dataset of 892 SRs separated into six classes, one for each plant species: *A. tomentosum* (96), *C. brasiliensis* (346), *M. guianensis* (36), *M. rubiginosa* (195), *P. ramiflora* (50), and *P. torta* (169).

Our strategy to evaluate vegetation indices in the context of plant identification relies on assessing the similarity among SRs associated with individuals of a same species. The similarity between two SRs is computed as a function of the distance between the time series obtained by computing a vegetation index for all days from our time period of study. A vegetation index is better than another if it ranks more SRs belonging to the same ROI of a query SR at the first positions. The objective of this experiment is to evaluate the ability of different vegetation indices in capturing leaf change patterns able to characterize the behavior of plant species or functional groups.

For each SR, we have computed 39 different time series by considering the image collection hours in each day (13 hours: from 6:00 to 18:00 h) and the normalized RGBcc indices (3 components: \( r_{cc}, g_{cc}, \) and \( b_{cc} \)), which are presented in Section 3. The distance function used for time series comparison is the Manhattan (\( L_1 \)) distance, as suggested by Conti et al. [2014]. In this way, we created 39 different descriptors to be used as input in our approach for vegetation index discovery. It is important to emphasize that any vegetation index could be used for time series extraction.

To create complex vegetation indices, a population of GP individuals were evolved along many generations using the GP framework (Section 4.2). Tournament selection was used to select GP individuals for applying genetic operators.

The fitness function adopted was FFP4 [Fan et al., 2004], which is defined as:

$$Fitness = \sum_{i=1}^{N} r(d_i) \times k_b \times k_g,$$

where \( r(d) \in [0, 1] \) is the relevance score associated to the returned results of a query, it being 1 if they are relevant and 0 otherwise; and \( k_b \) and \( k_g \) are two scaling factors adjusted to 7 and 0.982, respectively.

This fitness function follows the principles of utility theory [Fishburn, 1988]. According to the utility theory, there is a utility function that assigns a utility value to each collection item. This function can be seen as a users preference function, based on which a utility value can be computed. Each item (an image region, as in our case) has a different value. In general, we assume the utility of a relevant image region decreases with its ranking order. “More formally, given a utility function \( U(x) \), and two ranks \( x_1, x_2 \), with \( x_1 < x_2 \), according to this assumption, we expect the following condition to hold: \( U(x_1) > U(x_2) \)” [Fan et al., 2004]. There are many possible functions that can be used to model this utility function satisfying the order-preserving condition given above [Fan et al., 2004]. The use of FFP4 is motivated by the good results observed in previous studies [da S. Torres et al., 2009; Fan et al., 2004], which modeled the use of GP in search problems as we do in this paper.

The GP-based vegetation indices were compared against with the normalized RGBcc indices calculated from digital images taken at noon. This baseline is a general standard adopted in ecological studies [Ahrends et al., 2009; Ide and Oguma, 2010, 2013; Richardson et al., 2009, 2007].

The effectiveness of each approach was assessed using Precision \( \times \) Recall curves [Baeza-Yates and Ribeiro-Neto, 1999].

---

**Figure 5:** Flowchart of an image retrieval system used in the evaluation protocol. The process is composed of offline and online steps. The offline steps comprise: (1) the encoding of leaf change patterns from labeled image regions into time series by computing different vegetation indices along multitemporal data; (2) the modeling of complex vegetation indices by combining those time series through a genetic programming framework (Section 4.2); and (3) the indexing of the combined time series in a data repository. The online steps consists of a query search, which are composed of the following steps: (4) the selection of a query pattern, which is an image area along the multitemporal images; (5) the extraction of the query’s time series by computing the vegetation indices; (6) the combination of the time series by using the discovered vegetation indices; (7) the search computation by similarity; and (8-10) the final similarity ranking, including all patterns learned at the offline stages.
Precision is the ratio of the number of relevant SRs retrieved to the total number of irrelevant and relevant SRs retrieved. Recall is the ratio of the number of relevant SRs retrieved to the total number of relevant SRs in the database. Here, a given SR is considered as relevant only if it belongs to the same ROI of a query SR.

However, there is a trade-off between Precision and Recall, i.e., increasing Recall may decrease Precision and vice versa. For this reason, we also report the results using unique-value measurements: Mean Average Precision (MAP), which is the mean of the precision scores obtained at the ranks of each relevant SR, and Precision at 5 (P@5), which is the average precision after 5 SRs are returned. Such metrics combine both Precision and Recall into a single measure, making the comparison easier. MAP and P@5 are standard measures for evaluating the performance of image retrieval systems. We have employed those measures mainly because the evaluation of the GP-based obtained indexes was modeled as an image retrieval task. MAP can be seen as the area under the Precision-Recall curve. It works as an overall effectiveness index for image retrieval system evaluation. P@5 measures the relevance of the first 5 results returned by a system. It is an important indicator of whether an image retrieval system is successful or not, as the best group of results perceived by the system is always presented first. Other metrics, such as ROC-AUC, could be used if we model the evaluation of GP as a supervised classification problem. Anyway, it does not affect our general conclusions since there are a high correspondence between ROC-AUC and MAP (Davis and Goadrich, 2006).

A 5-fold cross-validation has been used to ensure statistically sound results. This procedure is performed as follows. Initially, we randomly partition the input dataset into 5 equal size subsets, called folds. Then, a GP-based vegetation index is learned using 4 folds and tested in the fold left out. This process is repeated 5 times, each time leaving a different fold aside as test set. The reported results refer to the average scores and the 95% confidence intervals obtained by the effectiveness measures for the evaluated vegetation indices, which were computed based on the mean and standard deviation of each fold.

### 6. Experimental Results

The GP framework requires five parameters: the size of population $N_p$, the number of generations $N_{gen}$, and the genetic operators rates (reproduction $r$, mutation $m$, and crossover $c$). The parameter $r$ was set to 0.05, as suggested by da S. Torres et al. (2009). We conducted initial experiments, aiming at determining the best values for the remaining parameters, based on results reported by da S. Torres et al. (2009). First, a population of 100 GP individuals were evolved along 10 generations considering $m = \{0.05, 0.1, 0.2\}$ and $c = \{0.2, 0.5, 0.8\}$. The higher effectiveness scores were obtained with $m = 0.1$ and $c = 0.8$, therefore these values were used for all other experiments.

Then, we performed a full factorial design to explore the parameter space of GP with respect to the population size $N_p$ and the generation size $N_{gen}$. The factorial design reveals the relative importance of each parameter, including cross-effects. The analysis of this experiment shows that the changes in the parameter $N_p$ do not cause a great impact on the results. The best scores for both measures were obtained with $N_p = 300$ and, therefore, we choose to report the results for this value. On the other hand, the effectiveness of GP is higher dependent on the choice of the parameter $N_{gen}$. In Figure 6 we show the MAP and P@10 scores achieved by the GP framework as the parameter $N_{gen}$ increases. We can see that, as more evolved are the GP individuals (larger $N_{gen}$), higher effectiveness scores are obtained, until reach an equilibrium, that is, a point from which the improvements are not statistically significant. To determine this point, we examined different values (from 5 to 200) to the low level setting for the parameter $N_{gen}$ and repeated the factorial ANOVA test until the null hypothesis was accepted, indicating that the effects of the evaluated parameters are not significant. According to this experiment, the results for $N_{gen} > 30$ are similar to those observed when $N_{gen} = 200$. For this reason, we used $N_{gen} = 30$.

Table I summarizes the results of the factorial analysis of variance (ANOVA). We separately analyzed the influence of the parameters $N_p$ and $N_{gen}$ for both measures (MAP and P@5). The percentages indicate the relative contribution of each parameter to the observed variation in each analyzed response. For instance, a simple change in the parameter $N_{gen}$ explains 38.72% of the variation observed in the P@5 measure. The models p-values indicate high statistic significance. The cross-effects were found not to be statistically significant.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Levels</th>
<th>MAP</th>
<th>P@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_p$</td>
<td>100</td>
<td>300</td>
<td>35.66%</td>
</tr>
<tr>
<td>$N_{gen}$</td>
<td>5, 30</td>
<td>200</td>
<td>38.72%</td>
</tr>
<tr>
<td>Model p-values:</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td></td>
</tr>
</tbody>
</table>

The best scores for both measures were obtained with $N_p = 300$ and, therefore, we choose to report the results for this value.

![Figure 6: MAP and P@5 scores at each GP generation step.](image)

Table I presents the best GP individuals (i.e., the GP-based vegetation indices) found by the GP framework in each of the

---

[1] Bimbo, 1999. The GP framework requires five parameters: the size of population $N_p$, the number of generations $N_{gen}$, and the genetic operators rates (reproduction $r$, mutation $m$, and crossover $c$). The parameter $r$ was set to 0.05, as suggested by da S. Torres et al. (2009). We conducted initial experiments, aiming at determining the best values for the remaining parameters, based on results reported by da S. Torres et al. (2009). First, a population of 100 GP individuals were evolved along 10 generations considering $m = \{0.05, 0.1, 0.2\}$ and $c = \{0.2, 0.5, 0.8\}$. The higher effectiveness scores were obtained with $m = 0.1$ and $c = 0.8$, therefore these values were used for all other experiments.

Then, we performed a full factorial design to explore the parameter space of GP with respect to the population size $N_p$ and the generation size $N_{gen}$. The factorial design reveals the relative importance of each parameter, including cross-effects. The analysis of this experiment shows that the changes in the parameter $N_p$ do not cause a great impact on the results. The best scores for both measures were obtained with $N_p = 300$ and, therefore, we choose to report the results for this value. On the other hand, the effectiveness of GP is higher dependent on the choice of the parameter $N_{gen}$. In Figure 6 we show the MAP and P@5 scores achieved by the GP framework as the parameter $N_{gen}$ increases. We can see that, as more evolved are the GP individuals (larger $N_{gen}$), higher effectiveness scores are obtained, until reach an equilibrium, that is, a point from which the improvements are not statistically significant. To determine this point, we examined different values (from 5 to 200) to the low level setting for the parameter $N_{gen}$ and repeated the factorial ANOVA test until the null hypothesis was accepted, indicating that the effects of the evaluated parameters are not significant. According to this experiment, the results for $N_{gen} > 30$ are similar to those observed when $N_{gen} = 200$. For this reason, we used $N_{gen} = 30$.

Table I summarizes the results of the factorial analysis of variance (ANOVA). We separately analyzed the influence of the parameters $N_p$ and $N_{gen}$ for both measures (MAP and P@5). The percentages indicate the relative contribution of each parameter to the observed variation in each analyzed response. For instance, a simple change in the parameter $N_{gen}$ explains 38.72% of the variation observed in the P@5 measure. The models p-values indicate high statistic significance. The cross-effects were found not to be statistically significant.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Levels</th>
<th>MAP</th>
<th>P@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_p$</td>
<td>100</td>
<td>300</td>
<td>35.66%</td>
</tr>
<tr>
<td>$N_{gen}$</td>
<td>5, 30</td>
<td>200</td>
<td>38.72%</td>
</tr>
<tr>
<td>Model p-values:</td>
<td>&lt; 0.05</td>
<td>&lt; 0.05</td>
<td></td>
</tr>
</tbody>
</table>

The best scores for both measures were obtained with $N_p = 300$ and, therefore, we choose to report the results for this value.

![Figure 6: MAP and P@5 scores at each GP generation step.](image)

Table I presents the best GP individuals (i.e., the GP-based vegetation indices) found by the GP framework in each of the
folds. Notice that vegetation indices related to different color channels were combined in a complex fashion, revealing a possible interaction between different phenology events ($r_{cc}$ is related to leaf senescence, whereas $g_{cc}$ is associated to leaf flushing). Also, observe that the extreme hours (morning and afternoon) were the most chosen to compose the GP-based vegetation indices. It indicates that early and late hours are better to distinguish the plant species. This finding disagrees with the general suggestion of analyzing digital images taken at midday hours reported in ecological studies (Ahrends et al., 2009; Ide and Oguma, 2010, 2013; Richardson et al., 2009, 2007).

Table 2: The best GP individuals found in each of the folds.

<table>
<thead>
<tr>
<th>Fold</th>
<th>The GP-based Vegetation Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(r_7 + (((b_7 * (r_{11} + r_{17}))/((b_{12} + r_6)) * (((0.52 * (r_{16} * (r_{15} + (r_{18} + b_{14}) - (r_6 + r_{13})))) * (g_7 + g_8))/b_6)) + ((r_5 + r_{15}) * (b_9 + r_6)) * ((g_{12} + r_3) * r_{10})) * (r_{18})$</td>
</tr>
<tr>
<td>2</td>
<td>$(0.64 * ((b_6 + (g_{16} + r_{18})) + (r_6) / p_{13})) + (((r_7 + r_{18}) + ((b_6 + ((g_9 + r_{11}) + (r_{17} + b_9)) - r_{10}))) + (g_9 + (r_{14} * r_3) + (((0.59 + r_{12})) * (g_{10} + b_{14}) * g_7)) * r_{16})) * (r_{17})$</td>
</tr>
<tr>
<td>3</td>
<td>$(((r_6 + r_{17}) + (((((r_{18} + b_{17}) + (r_{15} + r_{18})) / g_{14}) + (r/2)))) + (((0.68 + r_{14}) * r_{18})) + (((2.9 * ((b_9 + b_6)) * b_7) + (r_{13} + (g_{14} + b_{10})))) + (b_9 + r_{10})) + (g_8) / (r_{12})) / (g_7 + b_{15})$</td>
</tr>
<tr>
<td>4</td>
<td>$(((0.45 + ((b_6 + r_7) + r_6))) + ((r_3 + (b_{18} + r_{16}))) * (b_{15}) + (((r_{18} + g_8)) + (r_{18} / (r_{11} + b_{13}))) + (r_{18}) + (b_7 + (g_{10} + g_{11}) + (g_8 + b_{10}) + ((g_{13} + r_9) + g_{11}))))) * r_{17}$</td>
</tr>
<tr>
<td>5</td>
<td>$(((r_{18} + r_7) + (((0.52 * r_6)))) + (((0.11 + b_9)) + ((0.34 * (g_8 + (g_{10} + g_7))) + (g_{15} + b_{17}) + ((r_{10} / b_{12}))) + (0.62 + (((g_{16} / (0.11 * (g_{12} + r_{13} + g_7))) * r_9)) + b_{16} * r_8) * ((r_{14} + r_{13} + r_{10})) + b_9))$</td>
</tr>
</tbody>
</table>

Figure 7 presents the Precision × Recall curves obtained by the evaluated approaches.

Figure 8 presents MAP and P@5 scores obtained by each of the evaluated approaches. In general, those vegetation indices are designed to capture different leaf change patterns. On the other hand, their combination is promising, yielding the best results. Notice the superiority of the GP-based vegetation indices when compared with the normalized RGBcc indices.

In Figure 8 we compare the GP-based vegetation indices and the normalized RGBcc indices with respect to the MAP and P@5 measures, respectively. MAP is a good indication of the effectiveness considering all positions of obtained ranked lists. P@5, in turn, focuses on the effectiveness of the methods considering only the first positions of the ranked lists. Again, similar results to those observed for the Precision × Recall curves were obtained.

Paired t-tests were performed to verify the statistical significance of those results. For that, the confidence intervals for the differences between paired means of each ROI were computed to compare every pair of approaches. If the confidence interval includes zero, the difference is not significant at that confidence.
level. If the confidence interval does not include zero, then the sign of the difference indicates which alternative is better.

Table 3 presents the 95% confidence intervals of the differences between the normalized RGBcc indices and the GP-based vegetation indices for the MAP and P@5 measures, respectively. Since the confidence intervals do not include zero in any case, the results confirm that the GP-based vegetation indices outperform the normalized RGBcc indices for both measures.

Table 3: Differences between MAP and P@5 scores of the different approaches.

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP min.</th>
<th>MAP max.</th>
<th>P@5 min.</th>
<th>P@5 max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP - r_{cc}</td>
<td>0.046</td>
<td>0.004</td>
<td>0.165</td>
<td>0.052</td>
</tr>
<tr>
<td>GP - g_{cc}</td>
<td>0.157</td>
<td>0.047</td>
<td>0.272</td>
<td>0.115</td>
</tr>
<tr>
<td>GP - b_{cc}</td>
<td>0.089</td>
<td>0.019</td>
<td>0.206</td>
<td>0.101</td>
</tr>
</tbody>
</table>

In Figure 9 we compare the individual scores obtained for each ROI in terms of the MAP and P@5 measures, respectively. Observe the differences in responsiveness of the normalized RGBcc indices with respect to each of the species individually. Those vegetation indices reflect different phenology events: the changes in the r_{cc} index are related to leaf senescence whereas the g_{cc} index variations are associated to leaf flushing. Therefore, this behavior reflects the contrasting leaf phenology of the plant species (Alberton et al., 2014): in an evergreen species, the leaf senescence is a continuous process and changes in the r_{cc} index are more subtle over time; in contrast, in a deciduous species, the change in the g_{cc} index reflects the rapid leaf senescence and the flush of new leaves. This phenomenon evinces a clear indication that the combination of those vegetation indices may lead to better results. Notwithstanding, the GP-based vegetation indices outperform the normalized RGBcc indices for characterizing both types of species.

7. Conclusions

Vegetation indices play an important role in characterizing leaf change patterns of plant species for near surface remote phenology studies. This work presented and discussed a genetic programming (GP) framework for deriving appropriate vegetation indices based on learning phenological patterns from vegetation digital images.

A rigorous comparative analysis of vegetation indices, as well as their possible combinations created by our approach, has been conducted in our experiments. Results obtained with our proposed index applied to about 2,700 images taken from a tropical cerrado-savanna vegetation demonstrate that vegetation indices may provide different and complementary information that can be explored to improve their accuracy on characterizing plant species.

Directions for future work include the evaluation of other vegetation indices for phenology analysis, an extensive study on the influence of parameters employed in the GP framework in order to determine which configuration produces the best results, the use of other learning-to-rank methods for combining vegetation indices (e.g., hierarchical clustering [Rocha et al., 2008]), as well as perform a rigorous analysis of the time series generated by the GP in order to verify if they encode phenological patterns which are also validated by the on-the-ground phenology [Alberton et al., 2014]. Finally, we suggest a detailed investigation of the GP effects on other applications in earth observation and remote sensing.

Acknowledgments

This research was supported by the São Paulo Research Foundation FAPESP and Microsoft Research Virtual Institute (grants #2010/52113-5, #2013/50155-0, and #2013/50169-1). BA received a master scholarship from CAPES and a doctoral fellowship from FAPESP (grant #2014/00215-0); LPCM and RST receive a Productivity Research Fellowship from CNPq (grants 306243/2010-5 and 306587/2009-2). Also, we have been benefited from funds of CAPES, CNPq (grant 449638/2014-6), FAPESP (grants #2007/52015-0, #2007/59779-6, #2009/18438-7, #2010/51307-0), and