Measuring phenology uncertainty with large scale image processing

Guilherme Rezende Alles, João L D Comba, Jean-Marc Vincent, Shin Nagai, Lucas Mello Schnorr

Measuring Phenology Uncertainty with Large Scale Image Processing*

Guilherme Rezende Alles[a], João L. D. Comba[a], Jean-Marc Vincent[b], Shin Nagai[c] and Lucas Mello Schnorr[a,*]

[a]*Institute of Informatics, Federal University of Rio Grande do Sul – UFRGS*
91501-970, Porto Alegre, RS – Brazil
[b]*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG*
F-38000 Grenoble – France
[c]*Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology,*
3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa, 236-0001 Japan

ARTICLE INFO

Keywords:
Phenology analysis
Parallel Workflow
Phenological visualization
Mathematical modeling
Uncertainty Quantification

ABSTRACT

One standard method to capture data for phenological studies is with digital cameras, taking periodic pictures of vegetation. The large volume of digital images introduces the opportunity to enrich these studies by incorporating big data techniques. The new challenges, then, are to efficiently process large datasets and produce insightful information by controlling noise and variability. On these grounds, the contributions of this paper are the following. (a) A histogram-based visualization for large scale phenological data. (b) Phenological metrics based on the HSV color space, that enhance such histogram-based visualization. (c) A mathematical model to tackle the natural variability and uncertainty of phenological images. (d) The implementation of a parallel workflow to process a large amount of collected data efficiently. We validate these contributions with datasets taken from the Phenological Eyes Network (PEN), demonstrating the effectiveness of our approach. The experiments presented here are reproducible with the provided companion material.

*This research was partially supported by the FAPERGS Research Foundation (grants GreenCloud 16/488-9 and MultiGPU 16/354-8), the National Council for Scientific and Technological Development (CNPq) (grant 447311/2014-0), the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES) - Finance Code 001, the CAPES/Brafitec EcoSud 182/15, and the CAPES/Cofecub 899/18 projects.

*Principal corresponding author
gralles@inf.ufrgs.br (Guilherme Rezende Alles); comba@inf.ufrgs.br (João L. D. Comba); jean-marc.vincent@imag.fr (. Jean-Marc Vincent); nagais@jamstec.go.jp (Shin Nagai); schnorr@inf.ufrgs.br (Lucas Mello Schnorr)
ORCID(s):

1. Introduction

Plant phenology consists of the study of cyclical events in the vegetation lifecycle and the correlation of these events to external factors such as climate change Alberton et al. (2017) Walther et al. (2002) Morisette et al. (2009). In this field, scientists build historical data that allows them to identify patterns in plant behavior. The goal varies from a behavioral comparison of many years of observation Chmielewski and Rötzer (2001) to a correlation between the dataset information against external data such as atmospheric gas exchange HE et al. (2009). Additionally, phenology provides means for long term monitoring of a vegetation site, particularly for vast forests. With phenological analysis, it may be possible to detect anomalies such as deforestation, fire, and flooding in a remote site, as well as monitor vegetation recovery Alberton et al. (2017).

Phenological studies use features captured from the target vegetation as data sources. Sensors capture relevant information in a periodical fashion. The features are then analyzed to extract metrics that summarize the state of vegetation at a specific point in time. Further studies take advantage of such metrics to track vegetation changes over time Graham et al. (2009). One example of this strategy is NASA's MODIS project Zhang et al. (2003), where two satellites capture data in various groups of wavelengths. The Normalized Difference Vegetation Index (NDVI) has been used in conjunction with MODIS data to detect the presence of vegetation and evaluate its behavior Pettorelli et al. (2005), as well as provide remote monitoring of vegetation sites Beck et al. (2006). Although efficient, these specialized sensors are expensive for data acquisition. NVDI requires, for instance, that images contain visible to infrared light. This cost makes NVDI inaccessible for small and medium scale studies.

The usage of digital camera imagery, such as the example of Phenocam Richardson (2018), is a more accessible alternative for the analysis of phenological data. Previous studies Yan et al. (2019); Richardson et al. (2009); Sonnentag et al. (2012); Graham et al. (2009) show that it is possible to extract relevant metrics from pictures, identifying vegetation and phenological phases. The three color components (red, green, and blue) of each pixel are frequently the origin of the phenological metrics. Examples of RGB-derived metrics include 2G-RGBi Bater et al. (2011); Ide and Oguma (2010), the Green (g_{cc}) and Red Chromatic Coordinates (r_{cc}), and the Excess Green (ExG) Sonnentag et al. (2012). Most analysis compute the average of the metrics, such as the following for g_{cc} and r_{cc}:

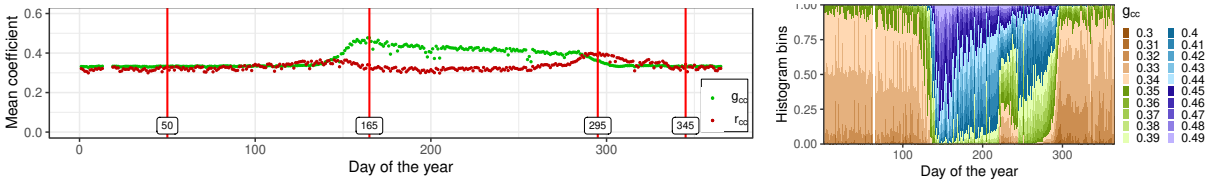
$$g_{cc}{}^{mean} = \frac{\sum_{i=0}^{n} \frac{G_i}{R_i+G_i+B_i}}{n} \tag{1}$$

$$r_{cc}{}^{mean} = \frac{\sum_{i=0}^{n} \frac{R_i}{R_i+G_i+B_i}}{n} \tag{2}$$

where R_i, G_i and B_i are the red, green and blue components (respectively) of a given pixel i, and n is the number of pixels in the input image (or the region of interest of the image).

Figure 1a depicts the mean g_{cc} and r_{cc} metrics of a deciduous broad-leaved forest in Japan as a function of time for one year. Such kind of plot enables one to correlate the high and low values to seasonal changes in the environment such as the green-up in spring and

senescence in the autumn Richardson et al. (2009). For example, we can see that there is a significant increase in the g_{cc} values during spring (green-up phase around day 165), while r_{cc} reaches its peak value in the autumn, with vegetation senescence (around day 295). The Figures in the bottom (1c, 1d, 1e, 1f) represent the original pictures taken at days 50, 165, 295, and 345 to illustrate typical views of those periods of the year.



(a) Average g_{cc} and r_{cc} values for TKY (2006).

(b) CPM for the g_{cc} metric.

(c) Day 50 (d) Day 165 (e) Day 295 (f) Day 345

Figure 1: Examples illustrating g_{cc} and r_{cc}-based metrics alongside four images.

Using average values of phenological metrics to depict vegetation behavior is very common Richardson et al. (2009); Sonnentag et al. (2012) but poses attenuation and accuracy problems. Averaging may attenuate essential nuances in the phenological metrics that could be vital for understanding the natural changes of phenophases. Accuracy problems may arise as shown in the example of Figure 1a, where we see that both g_{cc} and r_{cc} values present some variability from one day to the other. This variability happens because of noise in the source images, which can be the result of scene illumination shifts from one day to the other, shadows, obstruction of the camera lens, among others. Existing methods already tackle this variability problem. For example, Sonnentag et al. (2012) employs a sliding window and 90 percentile selection for noise reduction, but at the cost of temporal resolution. Another method includes the selection only of the midday image for each day Richardson et al. (2009), a frequent choice Leite et al. (2016), but disregarding the rest of the many potentially useful images captured per day.

Alternative methods based on histograms, such as PhenoVis' Chronological Percentage Map (CPM) Leite et al. (2016), employ stacked bar plots as shown in Figure 1b. Each vertical line has a stack of 20 rectangles whose height corresponds to the value of histogram bins, one for each interval of g_{cc} (represented by the colors). Since there is one stacked bar for each day, we can observe the changes of the histogram along time, a piece of information much more precise than the average g_{cc} (as in Figure 1a). The tool still lacks a mathematical model to smooth noise, and an appropriate palette to reflect colors from input images.

Contribution

We strive to build on the concept of CPMs to provide a framework for phenological analysis. Our solution tackles the visualization, accuracy, and reliability of the dataset information and computational scalability. (a) We incorporate improvements to the CPM visualization of PhenoVis, such as mapping colors from the input dataset to define the color palette of the final CPM. (b) We also propose three new phenology metrics based on the HSV color space while exploring how they depict vegetation phenophases using the CPM visualization. Regarding the reliability and accuracy of the dataset information, (c) we propose a mathematical model that derives metrics by assigning weights to a set of images in a given day, in order to smooth out the interference caused by external factors such as shadows, insects or condensation in the camera lens. Finally, (d) we propose a parallel strategy for our data analysis workflow, accelerating the analysis loop while incorporating multiple years and metrics at a time.

The paper is structured as follows. Section 2 details the contributions of our work: visualization, accuracy, and reliability of the dataset information and computational scalability. Section 3 presents the results of our method in the visualization, color representation, variability control, and uncertainty level measurement, while also discussing the scalability of our proposed workflow with performance analysis. Finally, Section 4 concludes the paper with a discussion about the obtained results.

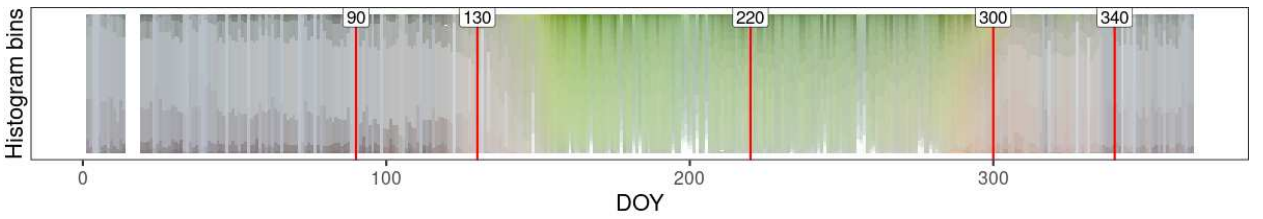
2. Materials and Methods

We describe in the following Subsections the methods we envisioned in our work to tackle uncertainty, improved visualization, and computational scalability.

2.1. New CPM color mapping enabling color-inspired phenology metrics

Chronological Percentage Maps from PhenoVis Leite et al. (2016) have a fixed color palette, being incapable of representing the same colors that are in the original pictures. This inflexibility may lead to misinterpretation of the resulting CPM. Improving the semantics of the color palette can increase the comprehension of the data Lin et al. (2013). We enhance the CPM to allow for an arbitrary color palette, enabling color definition for each bin in execution time. With this flexibility, the enhanced CPM depicts histograms with the same colors as the original pictures. Figure 2a depicts an example of the g_{cc} distribution through the year of 2006 from the Takayama deciduous broad-leaved forest site (TKY, 36º 08′ 46.2" N, 137º 25′ 23.2" E) Nagai et al. (2018). We see that colors used to depict each stacked histogram correspond directly to the colors from the pictures, as shown in the five pictures of days 90, 130, 220, 300, and 340 of that year. This mapping improves the reasoning process and the correlation to the input dataset by providing an apparent reference to the original data. As shown in Figure 2a, we can see five major regions of distinct colors along the span of the year: mostly grayed-out, from day 0 to day 120; transition phase around day 130; peak greenness at around day 220; vegetation senescence around day 300; and a cyclical return to the grayed-out state at day 340.

With color mapping flexibility, many choices exist to determine the colors for the CPM. As colors influence the perception of the data analyst, we propose four different phenology metrics for the CPM visualization in two groups: three HSV-based metrics,



(a) Enhanced color mapping for the CPM.

(b) Day 90 (c) Day 130 (d) Day 220 (e) Day 300 (f) Day 340

Figure 2: CPM of the g_{cc} (TKY, 2006), with a enhanced color mapping, matching the pictures.

which are inspired by the HSV (Hue-Saturation-Value) color space, and one g_{cc}-based metric. The HSV-based metrics generate a more precise histogram visualization, considering the color distribution of the original digital images. The g_{cc}-based metric, on the other hand, addresses the visualization issues that were pointed out in the original PhenoVis Leite et al. (2016) study, in which the g_{cc} distribution visualization lost the relationship with the source dataset by using a fixed color palette. For each metric, we generate histograms with the distribution of values for the CPM representation. We also remove irrelevant pixels, such as those depicting the sky or an observation tower, by masking images and selecting only the regions of interest.

HSV-based

The HSV (or HSL) color space is an alternative to representing colors without resorting to RGB values, which components are known to be deeply correlated Cheng et al. (2001); Pietikainen et al. (1996). Colors in the HSV color space have three components: hue (H), saturation (S), and value or luminance (V/L). By using the HSV color space to extract metrics from vegetation images, we expect to enhance visualization by presenting a more meaningful separation of color components, which should, in turn, allow for a better distinction of vegetation phenophases. The proposed HSV-based metrics are called `HSV_H`, `HSV_Mean` and `HSV_Mode`, as depicted in the three CPMs shown Figure 3 computed with daily images taken at noon, for simplicity. The differentiating factor among each metric is how they determine the colors to represent histogram bins.

The `HSV_H` histogram consists of the distribution of *hue* (H) values of the pixels in the image. Each histogram has 360 bins (the number of possible values for H), and the resulting CPM includes the sequence of calculated histograms. HSV components define the color of each histogram bin and are given by:

$$H_b = b \qquad S_b = 1 \qquad V_b = 1 \tag{3}$$

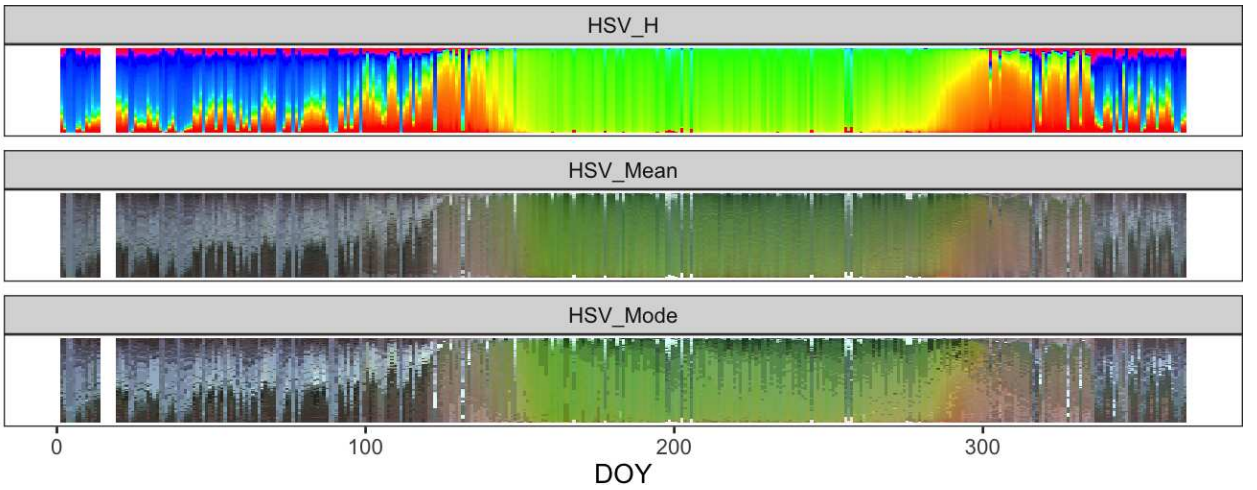


Figure 3: CPMs for the `HSV_H` (top), `HSV_Mean`, and `HSV_Mode` metrics (TKY, 2006).

where H_b, S_b and V_b are the HSV components that define the color of the b histogram bin. The `HSV_H` strategy generates vibrant colors. It enhances the perception of color distribution in the CPM because we set the maximum value (100%) for S and V, for every histogram bin. While this strategy depicts the vegetation green-up and senescence very clearly, the choices for S and V values sacrifice the dataset color representation since it employs saturated colors, making it unsuitable for any correlation with the input images. The other HSV-based metrics solve this problem.

The `HSV_Mean` metric computes the distribution of H values of images (as done previously), while also accumulating the S and V values of all the observations, grouped by bins. By accumulating the S and V components, we can then calculate the mean value of S and V of all the pixels classified in a given H bin. The `HSV_Mean` metric, thus, provides the same color distribution as the `HSV_H`, but with a more accurate color palette. The color of a histogram bin comes from three values. The bin's H component, the average of all the S values of pixels included in that bin, and the average of all the V values of pixels included in that bin. Mathematically, the color of the histogram bins defined by the `HSV_Mean` metric is given by:

$$H_b = b \qquad S_b = \frac{\sum_{i=0}^{N_b} S_{b_i}}{N_b} \qquad V_b = \frac{\sum_{i=0}^{N_b} V_{b_i}}{N_b} \tag{4}$$

where H_b, S_b and V_b are the HSV components of color for a given b bin; S_{b_i} and V_{b_i} are the S and V components of the i^{th} pixel classified in the b bin; and N_b is the total number of pixels categorized in the b bin.

The `HSV_Mode` metric follows a similar path of the previous metric since we also use the S and V values from the observations of each H bin to compute the resulting color. Instead of calculating the mean, however, we calculate the mode of the observations by building 10-bin histograms of S and V values for every bin. After building the histograms, we take the interval with most observations for S and V and use these components to

color the bin for the main histogram. The colors computed by `HSV_Mode` metric are given by:

$$H_b = b \qquad S_b = SMo_b \qquad V_b = VMo_b \tag{5}$$

where SMo_b and VMo_b are the mode of S and V of pixels classified in the b bin.

GCC-based

Our g_{cc}-based metric is an extension of the g_{cc} based visualization introduced by PhenoVis. This metric strives for using the Green chromatic coordinate as the base metric to identify phenophases, while also using information from the source images to keep the CPM representation resembling the original dataset. The histogram of the `Gcc_Mean` metric contains 100 bins. For every bin in the histogram, we accumulate the RGB components of all the pixels classified in that bin. The average of these RGB components defines the color assigned to the corresponding bin. As a consequence, alongside the g_{cc} histogram, we have the mean RGB components of pixels that contributed to each bin. We compute the RGB components of each bin's color as follows:

$$R_b = \frac{\sum_{i=0}^{N_b} R_{b_i}}{N_b} \qquad G_b = \frac{\sum_{i=0}^{N_b} G_{b_i}}{N_b} \qquad B_b = \frac{\sum_{i=0}^{N_b} B_{b_i}}{N_b} \tag{6}$$

where R_b, G_b and B_b are the R, G and B components which form the RGB color for a given b bin; R_{b_i}, G_{b_i} and B_{b_i} are the RGB components of the i^{th} pixel in the b bin; N_b is the number of pixels that were categorized in the b bin. Figure 4 shows a CPM calculated using the `Gcc_Mean` metric, considering daily images taken at noon.

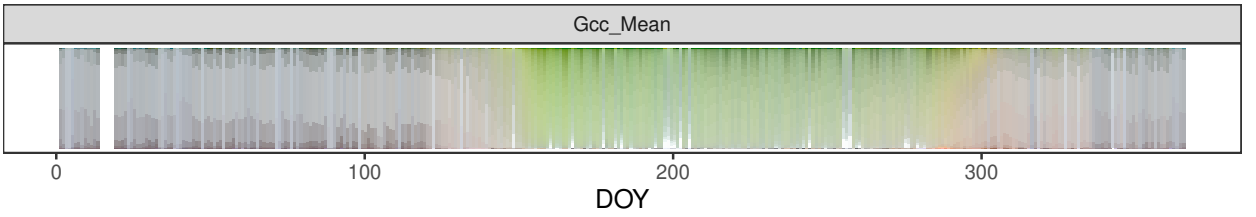


Figure 4: CPM for the `Gcc_Mean` metric in (TKY, 2006).

2.2. Mathematical modelling to control picture uncertainty

We propose a mathematical model that aims to reduce the uncertainty level by considering multiple images for a given period. Since different images may provide different contributions, because of the sunlight illumination and seasonal differences, each image in the period receives a user-configurable weight. We compute the representative histogram by applying those weights. The final CPM representation for that period is the one computed using this methodology.

User-configurable Weights. The weights definition is part of the analysis process. To accurately define weights, we need to evaluate the images that are available in the dataset and decide how much a given image should contribute to the weighted histogram, according to the capture hour and the symmetry of sunshine, for instance. For this study,

we define weights based on a Gaussian distribution around the mean, assigning the highest values to images taken between 12:00 and 13:00. We consider hours close to noon as the center of our weights distribution because this is the time at which we expect peak brightness and thus less variation from shadow exposure. These choices might change according to latitude and period of the year, but we fix them for simplicity. The resulting weighted histogram, which represents one day, is given by the sum of the product between every histogram and its respective weight:

$$H = \sum_t w_t . H_t \tag{7}$$

where H is the final weighted histogram (the resulting representative histogram for that time interval, plotted using the CPM method), H_t is the histogram calculated from an image taken at the hour t, and w_t is the weight assigned to the hour t according to our values inspired by the Gaussian distribution. H and each H_t are vectors of the same length, containing one place for each bin of the histogram.

Single-day Weighting. Figure 5 illustrates an example considering seven images taken from day 350 of 2016 of the Mt. Tsukuba dataset (MTK) and how our mathematical model acts when using the weights, as shown in Figure 10a. Figure 5a depicts the histograms of the original pictures, captured from 9:00 to 15:00. The images at the beginning of the day (9:00 and 10:00) are considerably different when compared to the rest. Orange and red hues are very pronounced in these cases but are much less present in images from hours 11:00 and 15:00. Such behavior indicates that during sunrise, the histograms can be considerably distorted, most likely by shadows in the input image. Our strategy, however, accounts for these variations by assigning very low weights to these images. The original pictures are subject to the defined weights to compute the resulting histogram, as shown in Figure 5b. By comparing this representative histogram to the ones from the original images, we can perceive that our approach tackles well those anomalies.

Multi-day Weighting. The histogram in Figure 5b is effective in smoothing out the trends present within each day. However, there can still exist high variability in the distribution of colors from one day to the next. This variability may come from conditions such as fog or rain, which can last several hours and thus affect many pictures in multiple days. In order to deal with such variations, we can extend the averaging process to calculate weighted CPMs of multiple days, as shown in Figure 10b. In this scenario, the day n, for which we are computing an averaged CPM, has a higher weight for its images compared to the weight of images from the previous two days ($n-1$ and $n-2$). We carry out the multi-day averaging process with a moving window.

The strategy of averaging images from multiple days with a moving window presents a tradeoff between smoothing and temporal size precision. One way to adjust this tradeoff is to change the size of the moving window: by increasing the window size, the CPM becomes smoother (less susceptible to high variability) and less precise. Decreasing the moving window size, on the other hand, enhances precision at the cost of creating a CPM with more color variation. Figure 6 shows CPMs that have been created from the same dataset using four window sizes: 2-day, 5-day, 7-day, and 10-day window. It is possible

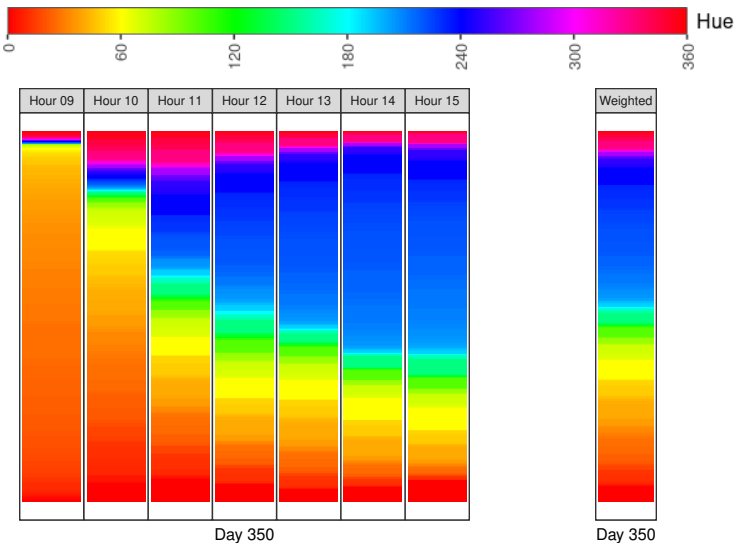


Figure 5: (a) The hourly histograms for the day 350; and (b) the histogram produced after applying weights to every hour. At the top, the hue values from the HSV color space and its respective color.

to observe the smoother plot as the window size increases. Albeit the lack of precision, a 10-day averaging (the bottom facet of Figure 6) provides an excellent insight into the general trend throughout the year.

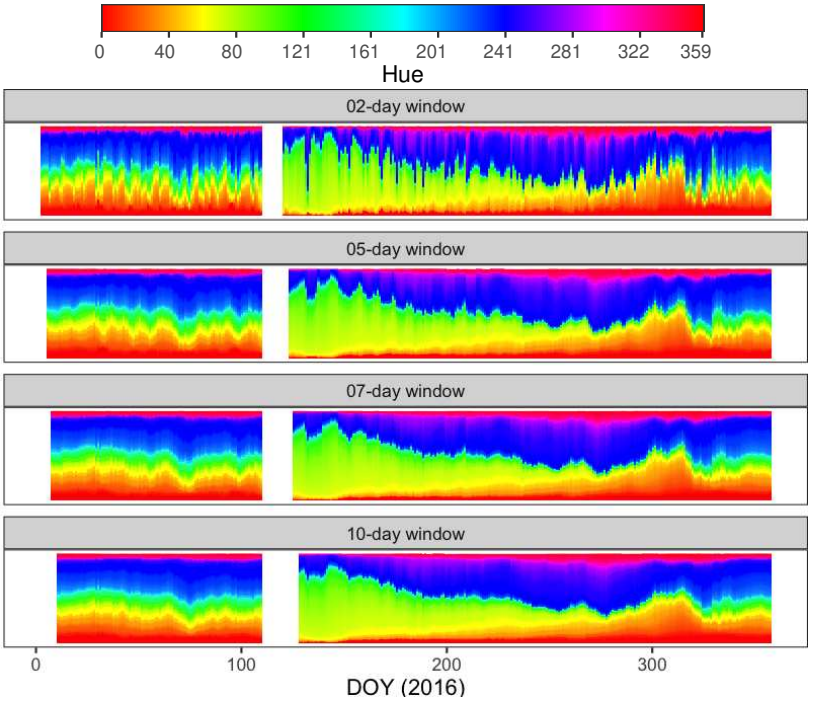


Figure 6: Averaged CPMs for MTK, 2016, with windows of size 2, 5, 7 and 10 days.

Uncertainty Level (of a histogram). By merging multiple histograms into one, we are effectively building a summary of a given day by analyzing different images. This combination is analogous to calculating the weighted average of many observations, which, in our case, is expected to reduce uncertainty by smoothing out abrupt variations in the distribution of colors. By averaging histograms, however, we are also introducing a new level of uncertainty in the representation. Since the calculated metrics are now a combination of multiple images, the final histogram for a given day no longer adequately represents one specific image from our source dataset. To quantify such uncertainty, we propose an **uncertainty level** metric to measure how well the weighted histogram represents all of the observations in a given day. The sum of all the differences between the weighted histogram and individual observations defines the uncertainty level of a weighted histogram. The differences are also weighted, as expressed by the following formula:

$$Q(H) = \sum_t w_t . d(H, H_t)^2 \tag{8}$$

where H is the weighted histogram, $Q(H)$ is the uncertainty level of the weighted histogram, w_t is the weight assigned to the hour t, $d(H, H_t)$ is the distance between the weighted histogram H, and the histogram of the image taken at the hour t. As before, H and each H_t are vectors of values, with one place for each bin of the histogram.

We employ the Earth Mover's Distance (EMD) Rubner et al. (2000) to compute $d(H, H_i)$. The EMD between two histograms X and Y defines the necessary effort to transform X into Y. In an iterative fashion, we compute the difference between subsequent bins, from EMD_0 to EMD_i for every i bin of X and Y histograms, as follows:

$$\begin{aligned} EMD_0 &= 0 \\ EMD_{i+1} &= X_i + EMD_i + Y_i \end{aligned} \tag{9}$$

where X and Y are vectors of values, with one placement for each bin of the histogram. When computing EMD, the beginning of the accumulation is a parameter, as it may change the final EMD value. In our work, we always start with the first bin: the H=0 bin when using the HSV-based metrics, or the bin identifier for the smaller g_{cc} value when using the g_{cc}-based metric. The bin order in the CPM justifies these choices. To ease the comparison between X and Y histograms, we need a single metric that represents the total distance. As it is familiar with the EMD metric, we compute the final EMD_{total} between two histograms by accumulating the absolute values of the differences between every bin i, as follows:

$$EMD_{total} = \sum_i |EMD_i| \tag{10}$$

Combining weighted histograms with their uncertainty levels enables us to indicate how well each weighted histogram represents the reality of each time interval it represents. Since the sum of differences gives the uncertainty, a value closer to zero indicates a more accurate representation. Figure 7 illustrates how such uncertainty level accompanies the CPM using the year 2016 of the MTK dataset. The uncertainty level appears as an

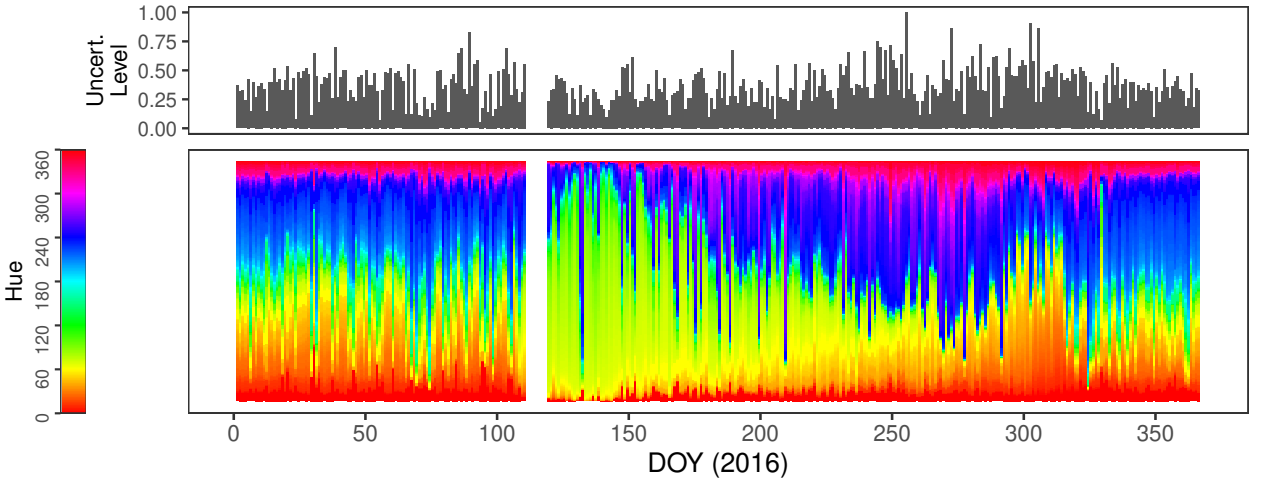


Figure 7: Uncertainty level (top) plotted alongside the CPM (bottom). This level indicates how uncertain each of the columns of the CPM is when the CPM aggregate many weighted pictures. At the left, we show the hue values from the HSV color space and its respective color.

indicator of how accurate the approximations are across the span of one year. The uncertainty plot appears above the CPM as a column chart. In this plot, every column indicates the uncertainty level of the matching histogram, represented in the same X coordinate in the CPM below. While the CPM of this Section considers the `HSV_H` metric, the averaging process itself is generic enough to be applied to any histogram. An example of applying the same process to the `GCC_Mean` metric is available in Section 3.3.2.

2.3. Two-Phase Workflow Implementation

The implementation aspect of our proposed phenological analysis process takes the form of a two-phase workflow. Figure 8 illustrates the general architecture of our implementation, detailed below. While Phase 1 must be executed in parallel since it is computationally expensive, Phase 2 runs on a laptop.

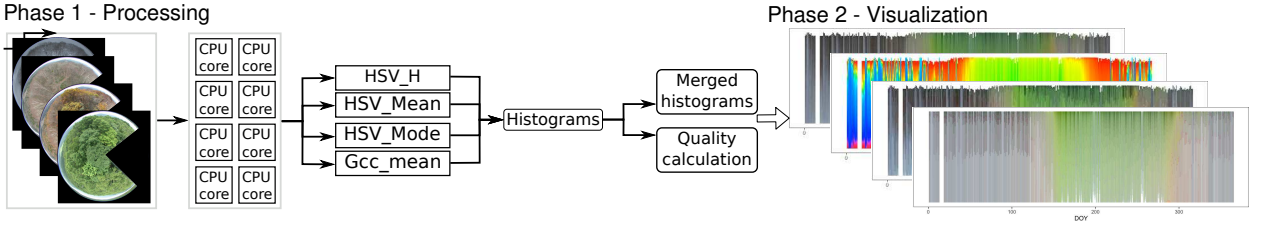


Figure 8: The proposed phenological analysis workflow

Phase 1. The core concept of the first phase is metrics extraction. It is done with a novel C++ implementation of PhenoVis called `phenovisr` which is an R package responsible for reading an input image and calculating the `HSV_H`, `HSV_Mean`, `HSV_Mode` and `Gcc_Mean` metrics. Optionally, it is also possible to provide a mask to `phenovisr`. The mask enables the selection of regions of interest in the input image. When a mask is available, `phenovisr` will ignore all the pixels that are covered by it. An R script

Table 1
Description of the deciduous broadleaf (DBF) and mixed forest datasets.

ID	Description	Type	Years	Image Res.	Images	Size	Pixels
TKY	Takayama (JAP)	DBF	2006–2012	2272x1704	20561	18GB	1221518
AHS	Alice Holt (UK)	DBF	2009–2017	2272x1704	55139	45GB	1166644
MTK	Mt. Tsukuba (JAP)	Mixed	2007–2017	2272x1704	71875	65GB	522560

equipped with the `parallel` package orchestrates the execution of `phenovisr`. It dispatches multiple executions of the metrics extraction routine in parallel. The return value from `phenovisr` is a data frame containing all of the information needed to build the CPMs. We keep this data frame in the local storage through a compressed `CSV` file.

Phase 2. The second phase consists of the data visualization, built using the R programming language. This part of the workflow expects an input file containing all of the necessary data to build the CPMs. The input contains the histograms for the HSV and g_{cc}-based metrics enriched with metadata, like the day of the year, time, dataset, and camera identifier for the processed image. In total, the input data format is a data frame expected to contain 460 lines for each image: 360 for the HSV histogram and 100 for the g_{cc} histogram. We used data manipulation with the `dplyr` package to extract the information needed for each plot, and the `ggplot` package to build figures. We rely on the flexibility of the scripting approach to provide the user with the freedom to adapt the resulting CPM as desired.

3. Results

We evaluate our methods with high-resolution pictures from the three datasets listed in Table 1 Nagai et al. (2018). We first evaluate the proposed phenology metrics and their CPM representation with the AHS dataset (Section 3.1). We assess the averaging strategy of weighted histogram using the MTK dataset (Section 3.2), followed by the uncertainty level evaluation with the TKY dataset (Section 3.3). Finally, we provide a computational scalability evaluation using the three datasets (Section 3.4).

3.1. Metrics and Color Mapping Evaluation

We analyze the enhanced CPM and the four metrics described in Section 2.1 against images from the Alice Holt (AHS) forest site. While the dataset has images from 2009 to 2017, we analyze only images fro 2012 to 2016 due to inconsistencies in camera position and orientation. We built the CPM visualization with images taken at noon, assuming that these provide lower noise levels due to peak sunlight to focus on the four metrics.

Figure 9 shows the yearly CPMs for each metric. We can see that the `HSV_H` metric enhances the perception of color distribution by increasing the saturation and luminance levels to 100%. This choice of values generates vibrant colors that indicate vegetation lifecycle periods throughout the year (marks A and B). However, the resulting representation harms the overall perception of the dataset since there is no clear relationship between the colors displayed in the CPM and the colors that are present in the original

images. Apart from elemental identification, analysis with this metric becomes reasonably limited. The `HSV_Mean` metric generates a color palette that resembles the original vegetation pictures, with some distinct aspects such as terrain (mark C). The averaging process, however, harms the color distinction between adjacent H bins. The alternative then is the `HSV_Mode` metric, which presents the same distribution with a different color palette. By using the mode of S and V between every H bin, this color palette effectively creates "steps" that make adjacent bin colors more visible than the `HSV_Mean` counterpart. Comparing `HSV_Mean` and `HSV_Mode` CPMs, we can distinguish unexpected data such as the one highlighted in mark D. This noise was due to condensation in the camera lens by manually inspecting the source images. Finally, the `Gcc_Mean` metric is unable to provide insightful conclusions. As foreseen in previous work Leite et al. (2016), different shades of green map to very similar g_{cc} values. As a consequence, different colors map to the same g_{cc} distribution bin, and since the colors of each g_{cc} bin define the CPM color palette, the colors can become easily distorted.

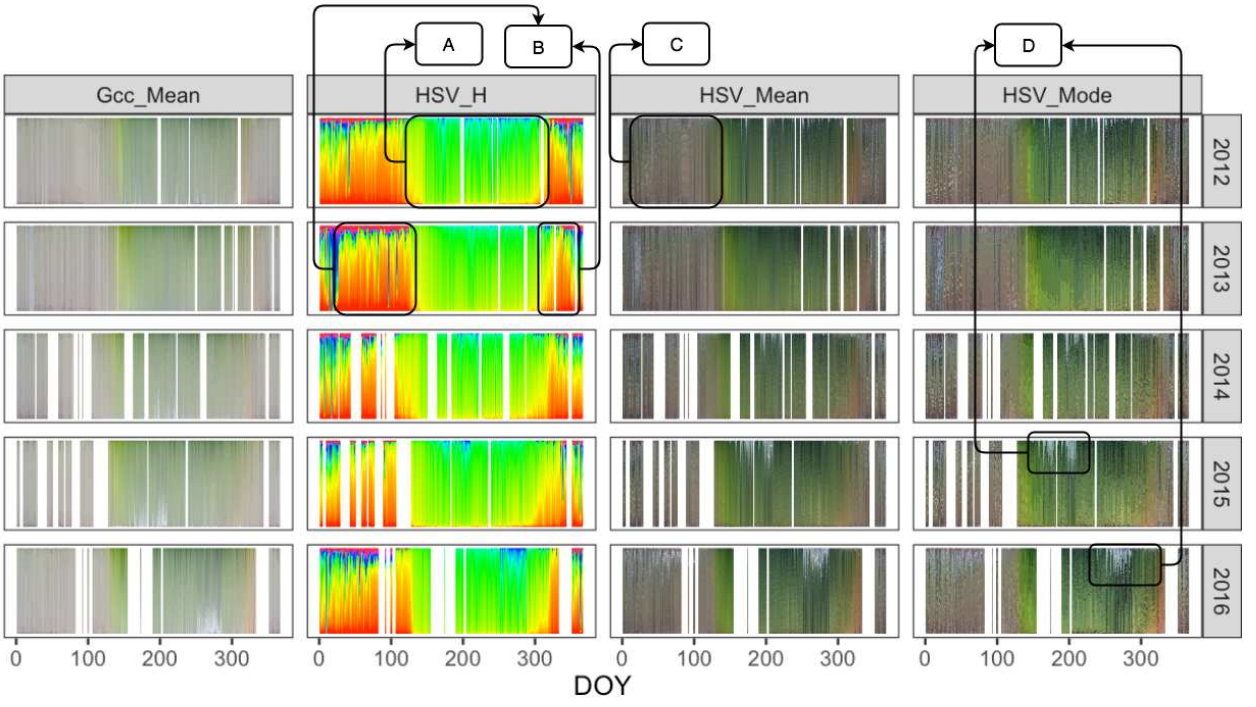
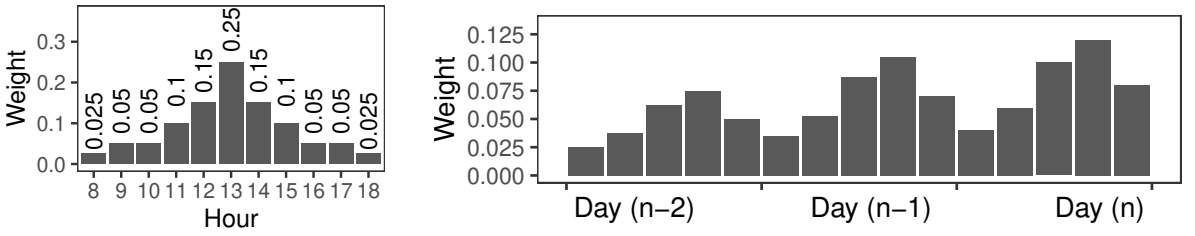


Figure 9: CPMs corresponding to four metrics (columns) of five years (lines) for the AHS dataset.

3.2. Weighted CPM Evaluation

We assess the method of building the weighted histogram by evaluating sets of images that become one. For this experiment, the dataset from Mt. Tsukuba (MTK) is suitable because it has a large number of images per day, from sunset to sunrise. Figure 10a depicts the weights of each hour applied each day for this dataset, while Figure 10b describes the 3-day weights of the experiment with the moving window across multiple days.



(a) Weights (Y axis) per hour (X) applied each day.

(b) Moving window used to average all histograms. For each day (n, $n-1$ and $n-2$), we take hours from 10 to 14.

Figure 10: Weights for the CPM evaluation using the MTK dataset using years from 2011 to 2016.

3.2.1. One-day averaging

Our first experiment consists of merging all the hourly images within a single day to generate a single distribution per day. Figure 11 shows the hourly `HSV_H` and `HSV_Mode` CPMs (facet columns) for the year 2016. These CPMs depict the variation in color throughout the day (facet lines). It is noticeable that colors are more distorted near the start and the end of the day (marks A and B), especially with `HSV_H` metric. For the start of the day, evaluating the color distribution in the `HSV_Mode` metric shows the predominance of orange and yellow hues (mark A.1). They indicate the occurrence of sunrise, which leads to noise due to sunlight exposure. The CPMs for the end of the day present some anomalies when analyzing the `HSV_H` metric, as shown in mark B.1. By analyzing the same period with the `HSV_Mode` metric, we can see that such anomalies are present because the pictures were taken at night (mark B.2). These plots also show that sunset happens earlier by the end of the year. After day 300, half of the pictures are from a night period. Marks C.1 and C.2 depict one specific scenario in which the camera lens faces the sun during sunset. The sunlight explains the high amount of orange and yellow hues in mark C.1.

Figure 12 shows the CPM computed by merging the weighted hourly data shown in Figure 11 using the weights from Figure 10a. Each merged CPM has the uncertainty level on top, which is the same for both since both metrics are HSV-based. When comparing to the original CPMs, we can see that our approach for increasing reliability is useful in producing a smoother plot. There is a high variability of color distribution, seen especially in the early and late hours of each day in Figure 11. The more significant weight applied for less noisy pictures (at noon, for instance) is supposed to compensate for the noise on early and late images. However, the high scores in the uncertainty level plot, also computed using the weights, indicate that the merged CPMs, in this case, have low quality because there is still too much dispersion in the original pictures. We can validate this claim by vertically comparing the histograms within mark D of Figure 11, which shows the high variability on images of different hours on the same day.

3.2.2. Multi-year data analysis

By applying a moving window across multiple days to generate an averaged CPM, we are favoring data summary and general trends instead of temporal precision. This trade-off can be justified when considering the data analysis of multiple years, in which the main objective is to observe the overall vegetation changes through long periods. Since

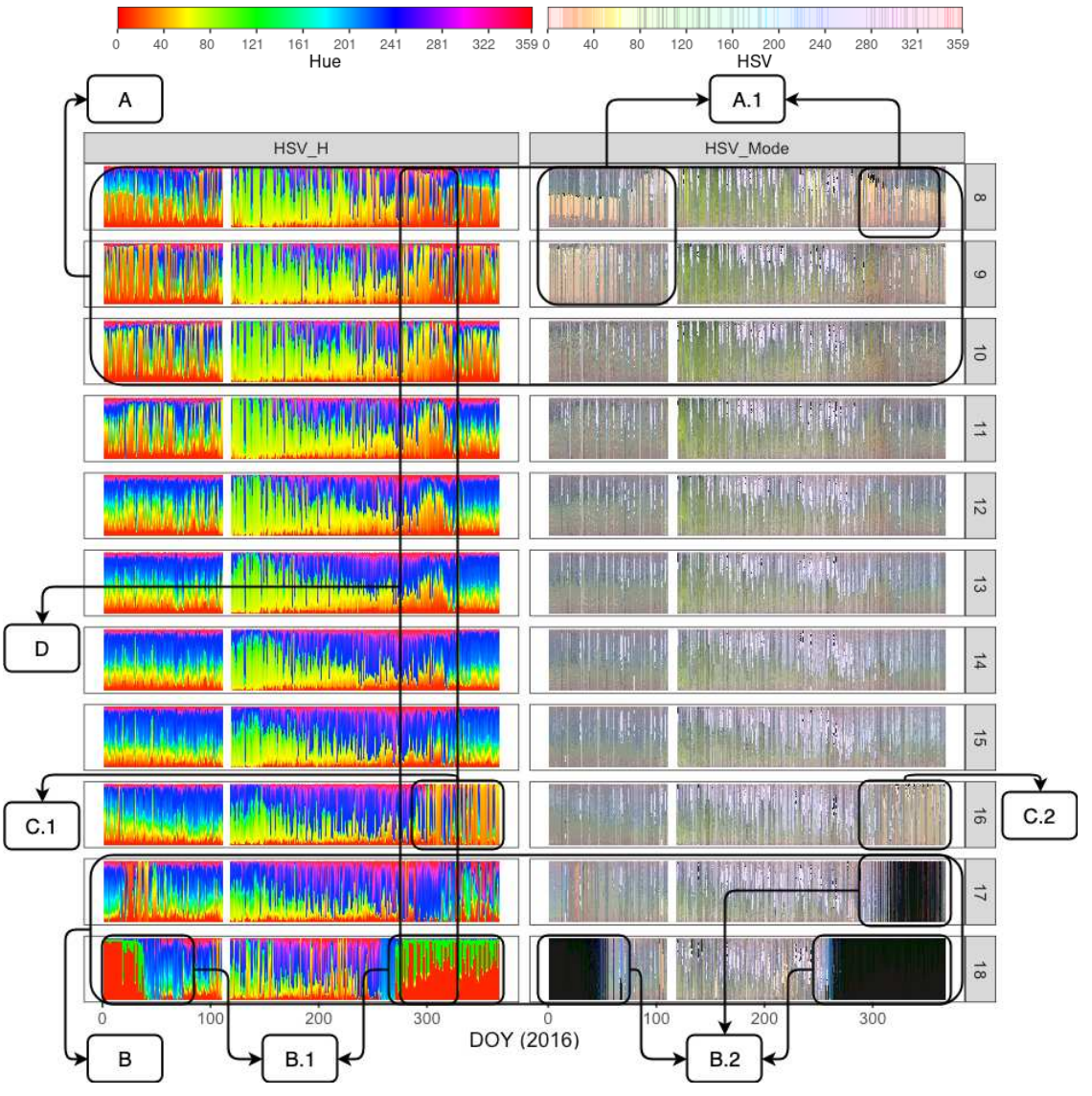


Figure 11: CPMs of the MTK dataset, grouped by hour (facet lines). At the top palettes, we show the available colors throughout each respective plot concerning the hue value.

we are comparing coarse-grained information (that is, the overall state of vegetation), temporal resolution is a secondary concern. Our focus is on the trend. To demonstrate this scenario, we applied the moving window multi-day averaging across multiple years in the MTK dataset. With the available data, we create CPMs from the years 2011 to 2016. The moving window used in this experiment is shown in Figure 10b.

Figure 13 shows the CPMs created by applying the moving window averaging across the years 2012 through 2016. The generated plot effectively depicts the vegetation state in each year, and the averaging across multiple days smooths the visualization. The multi-day averaged CPMs allows us to identify vegetation change and compare a set of days throughout many years. We illustrate some examples in the CPMs. The end of 2014 presented a few red hues in the color distribution when compared to the same period

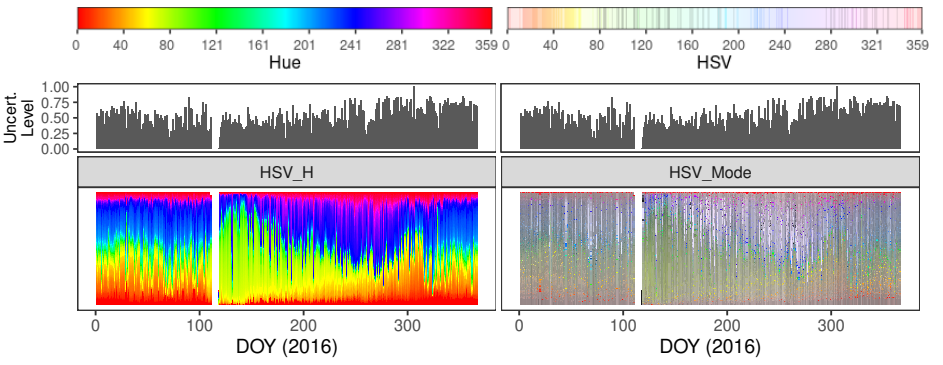


Figure 12: Merged `HSV_H` and `HSV_Mode` CPM for the MTK 2016, showing the uncertainty level.

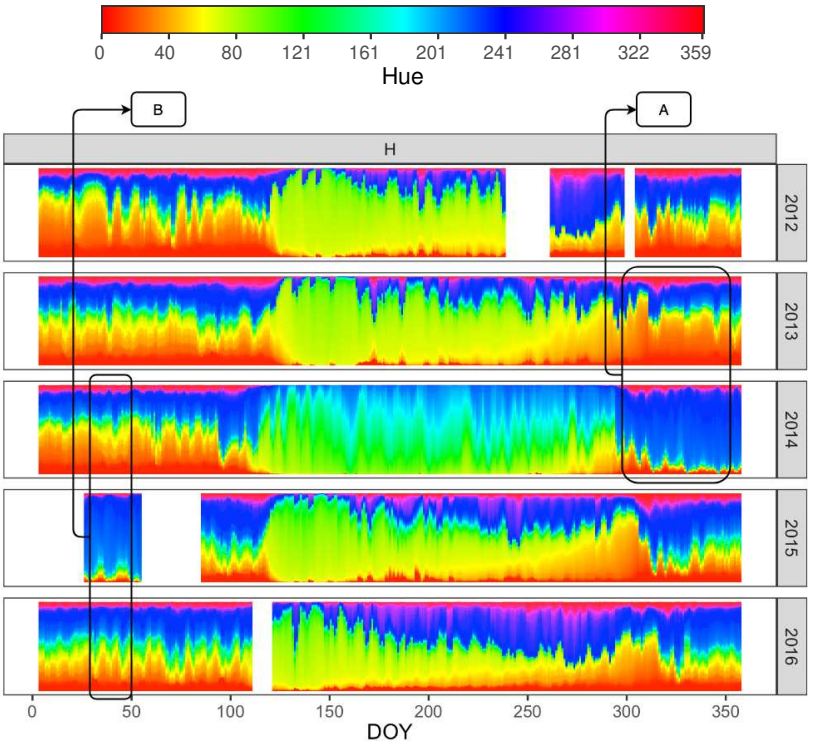


Figure 13: CPMs of the `HSV_H` metric by applying the moving window averaging, for five years.

in the year 2013 (mark A). One possible explanation for this distribution difference is the early leaf fall in 2014: for the same period, there are still some leaves in the canopy of the forest in 2013, while in 2014, leaf fall was already complete. Another example appears around DOY 50 of 2015, in which there is more predominance of blue hues when compared to the same period of 2014 and 2016 (mark B). A blizzard, absent in 2016 and less severe in 2014, covered the vegetation area around the camera, being the origin of the anomaly.

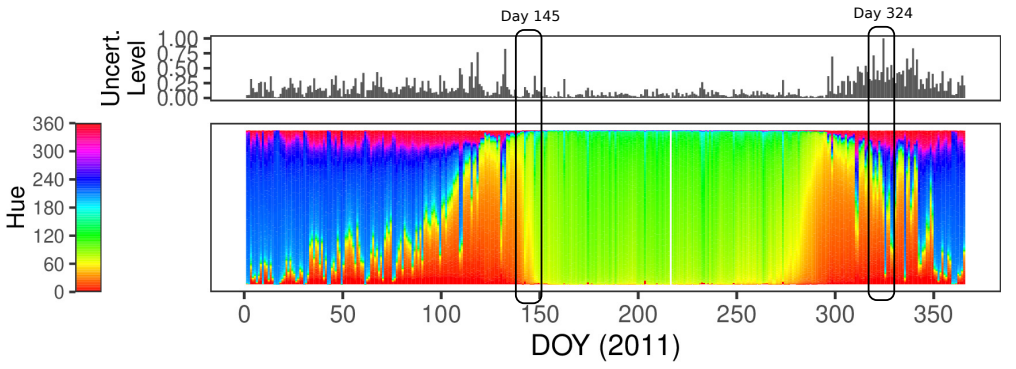


Figure 14: Weighted `HSV_H` CPM for 2011, with the uncertainty level of each day at the top.

3.3. Uncertainty Level Evaluation

To evaluate the distance between a weighted CPM against the one from original pictures, we select days with reported high and low levels of uncertainty and analyze the hourly images and CPMs for these days. Since the g_{cc}-based and HSV-based CPMs generate distinct distributions for the same image, there are then two possibilities for histograms to derive the uncertainty level: g_{cc} CPMs and HSV CPMs. We discuss the characteristics of each of them. The dataset from Takayama Flux Site (TKY) was the basis for the analysis, with images taken from the year of 2011. The merging process considers the hourly images within every day, as described in Section 3.2.1.

3.3.1. Uncertainty level based on HSV

The uncertainty level is plotted above the CPM in Figure 14. The plot presents a column chart with the EMD-based uncertainty level of each day. Based on the uncertainty level plot, we select two days – 145 and 324 – with low and high uncertainty levels for comparison.

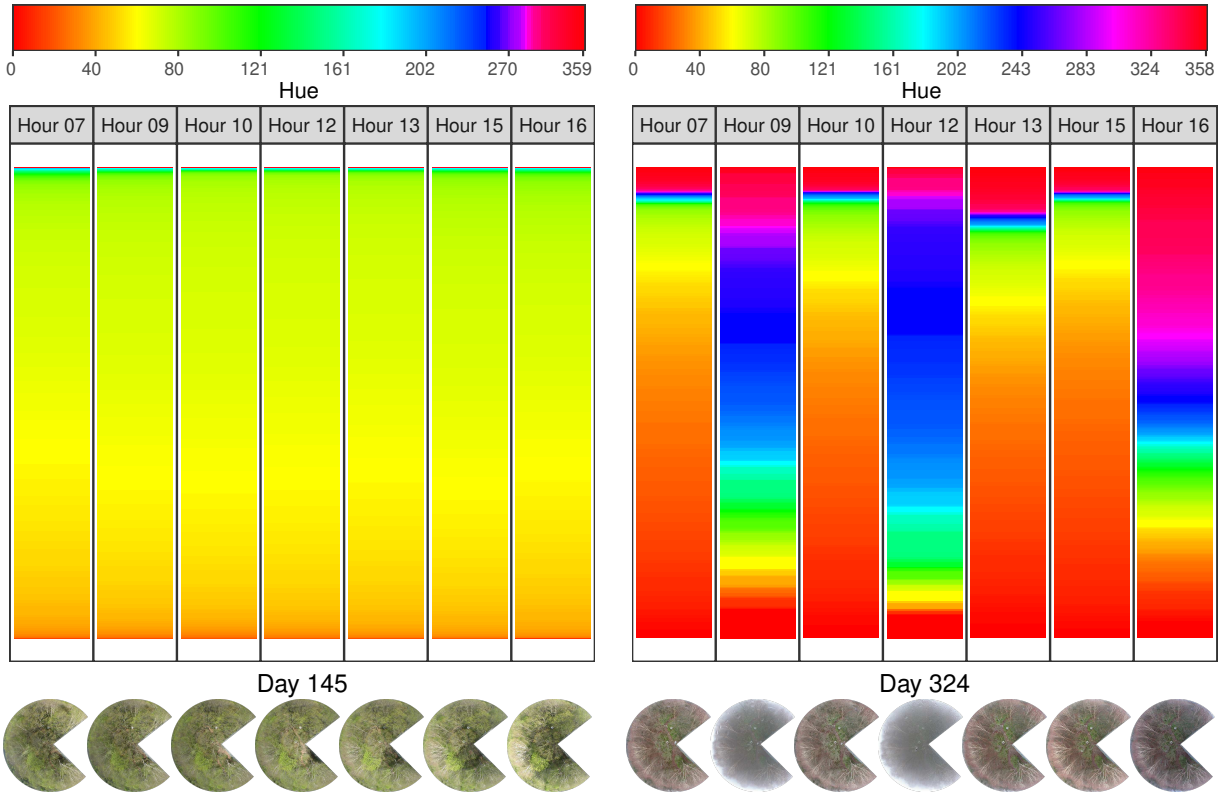
Figure 15a shows the hourly CPMs for day 145, for which we have identified a low uncertainty level. We notice that the hourly CPMs for this day are very similar, agreeing with the small variations in each of the corresponding pictures in the bottom. Such a combination reinforces the low uncertainty value computed for this day. Alternatively, Figure 15b shows the hourly CPMs for day 324, for which we have identified a high uncertainty level. Here, the CPMs are very different from hour to hour. The differences come by condensation in the camera lens, especially in pictures taken at hours 9 and 12, but also at 16. The high uncertainty value computed for this day captures these issues.

3.3.2. Uncertainty level based on $\mathbf{g}_{cc}$

Another possibility is to calculate the uncertainty level when using g_{cc} histograms. Figure 16 shows the weighted CPM for the `Gcc_Mean` metric, alongside the EMD-based uncertainty levels for the g_{cc} histograms. The calculated uncertainty level for g_{cc} is much more stable than the HSV counterpart (shown in Figure 14). Variability level is reasonably low across most of the year, and it increases considerably as g_{cc} levels become higher during vegetation green-up (between days 150 and 275, approximately). The increase happens because of the nature of g_{cc}, whose value correlates to the shades of green pixels

observed in the source images. Fewer shades of green in the pictures become histograms with g_{cc} distributions with very low variation.

Figure 17a shows the evolution of the weighted distribution of g_{cc} values across the



(a) CPMs and source images from DOY 145. (b) CPMs and source images from DOY 324.

Figure 15: Comparison of extreme cases of uncertainty levels.

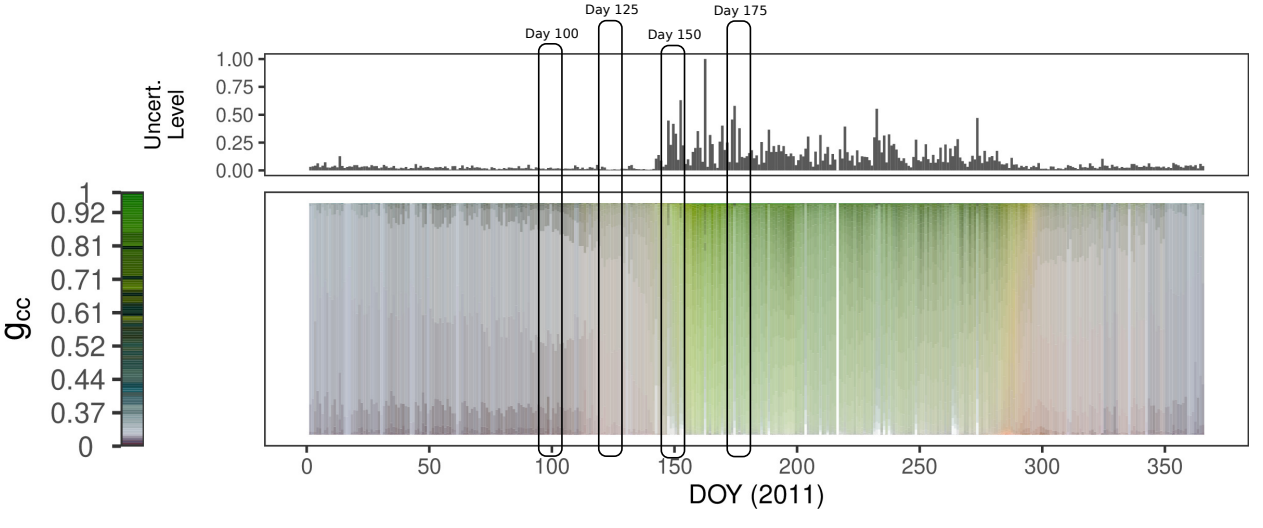
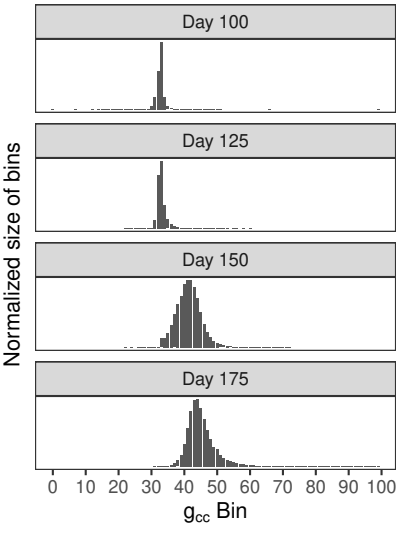


Figure 16: Weighted `Gcc_Mean` CPM for 2011, with the uncertainty level at the top.

start of the green-up phase to illustrate the contrast coming from the g_{cc} metric. Days 100 and 125, which presented low g_{cc} values and low uncertainty level, show a narrow histogram with the g_{cc} values being distributed across approximately ten bins (from 25 to 35). Days 150 and 175, on the other hand, presented higher g_{cc} values and higher uncertainty levels. The g_{cc} distribution for these days is much broader, which in turn allows for more variations between similar histograms.

We select DOY 162, with the highest uncertainty level for 2011, to demonstrate further the high sensitivity of the g_{cc}-based uncertainty level. Figure 17b depicts the hourly CPMs and corresponding pictures. We can see the images present some degree of variation due to condensation in the camera lens, which impacts the CPMs for hours 09 and 12. However, the color information is relatively preserved and contrasts the high uncertainty metric presented by our g_{cc} based calculation.



(a) Weighted histogram of g_{cc} with 100 bins for days 100, 125, 150 and 175 of 2011.

(b) CPM of the g_{cc} metric and corresponding pictures of DOY 162/2011.

Figure 17: Demonstration of the high sensitivity of the g_{cc}-based uncertainty level.

3.4. Metrics Extraction Performance Analysis

Table 1 lists the characteristics of the datasets of our performance analysis. The table's last column refers to the number of pixels considered for metrics extraction after the selection of the region of interest described in Section 2.1. For these tests, we executed the metrics extraction routine in one compute node of the PCAD (`http://gppd-hpc.`

inf.ufrgs.br/) from the Informatics Institute at UFRGS. This node is equipped with two Intel Xeon E5-2650 v3 processors (ten cores at 2.3GHz each), adding up to 20 CPU cores and 128GB of DDR4 RAM.

To measure the computational scalability, we designed an experiment to extract the execution time of our workflow with different configuration scenarios. Our experiment consisted of executing the histogram extraction multiple times, both sequentially and in parallel. For the parallel executions, the number of worker threads ranged from 4 to 40 in increments of 4. The decision of spawning more working threads than the number of cores in the testbed is two-fold. First, to account for simultaneous multithreading, which enables up to 40 computing threads in the node. Second, to measure whether a highly IO bound workflow (such as reading images from the disk) would benefit from an oversubscription environment. We replicate ten times each experimental configuration, in a random order, to address the variability of the experimental results.

Figure 18 shows the average execution time for Phase 1 (in the Y-axis) as a function of the number of workers (X-axis) and datasets (line colors). We observed a reasonably high execution time using the trivial, sequential implementation of the metrics extraction workflow. The execution time in this scenario ranges from approximately 51 minutes for the fastest dataset (TKY) to approximately 131 minutes for the slowest (AHS). We expect such variation in execution time because of the differences in workload size: the AHS dataset has almost three times more data when compared to the TKY dataset. Still considering the sequential implementation, the MTK dataset executes faster than AHS despite its larger size. The difference is due to lighter computational workload per image since the number of pixels considered for metrics extraction in the MTK dataset is roughly half as much as that of the AHS dataset, as detailed by the Pixels column at Table 1.

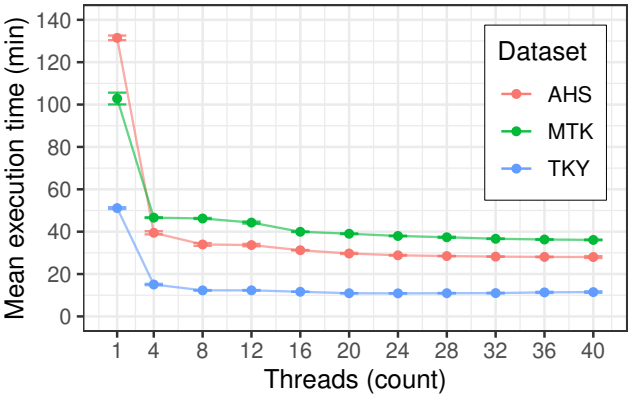


Figure 18: Average makespan (in minutes) of the metrics extraction process.

The parallel implementation significantly improves performance, decreasing the execution time to approximately 39, 46, and 15 minutes for the AHS, MTK, and TKY datasets. In this scenario, we noticed that our experiments with the AHS dataset presented better scaling when compared to the MTK dataset since the execution time for the former is lower when compared to the latter. This behavior indicates that the size of each computational task, indicated by the number of pixels considered per image for metrics extraction,

considerably influences scaling on parallel environments. Further increasing parallelism for up to 40 worker threads, however, continued to yield consistently lower performance benefits.

Although performance gains are noticeable, our experiments showed that we quickly hit diminishing returns. The difference in execution time by increasing the threads count to more than eight is reasonably small. On this matter, we attribute the diminishing returns to the highly IO bound nature of our workflow. Even though we can process data in a massively parallel fashion, we are still reading images from a spinning hard drive, which quickly becomes our bottleneck. One way to attenuate such a bottleneck is to explore parallel file systems for higher performance on data ingestion and investigate the distribution of subsets of data into multiple computing nodes.

4. Conclusion and Future Work

The phenological analysis is an essential aspect of biology study since it strives to understand the cyclical effects that are present in nature. When applied to vegetation, one of the main objectives of phenological studies is to try to correlate variations in plant lifecycle to external factors such as climate change.

To enhance the phenological analysis process using vegetation pictures, we created an analysis workflow that combines parallel processing for efficient metrics extraction with the visualization techniques needed to build Chronological Percentage Maps. Our proposed analysis incorporates three main aspects: color representation, reliability, and uncertainty measurement.

For color representation, we presented four metrics extracted from vegetation digital images: `HSV_H`, `HSV_Mean`, `HSV_Mode`, and `Gcc_Mean`. For each of these metrics, we extract both the histogram and associated metadata from the input dataset. The metadata associated with the histogram allows us to create a custom color palette, which in turn allows for the CPMs to have colors that resemble the original data. For the HSV based metrics, this approach improves visualization by enabling artifacts such as leaves, soil, snow, and the sky to be identified in the final CPM. For the g_{cc} based metric, however, the resulting CPM does not yield satisfactory results because many different colors can map to the same or very similar GCC values, which in turn makes the color representation for these CPMs not representative of the original data.

The reliability of color representations improves through merging histograms of subsequent images, assigning higher weights to those that are most relevant (such as images taken at noon, with peak sunlight). The design of the merging process accounts for abnormal variations on scene illumination, condensation, or obstruction of the camera lens or other anomalies. We also proposed a moving window strategy for merging histograms of subsequent days, to produce smoother CPMs that enhance data perception on noisy datasets. In our tests, both merging processes successfully generated less noisy CPMs by considering multiple image histograms with proper weights.

To consider the uncertainty introduced by the histograms merging process, which generates an approximation of the real observations, we used the Earth Mover's Distance (EMD). This distance indicates the distance between actual and computed histograms. By

plotting the EMD alongside each histogram in the merged CPM, we can effectively tell how close each approximation is to the real pictures. In our tests, the uncertainty level calculated from HSV based histograms was much more representative than its counterpart calculated from g_{cc} based histograms. We observed that g_{cc} based EMD tends to be extremely sensitive to histograms from images with high levels of green since the g_{cc} value is deeply dependent on the proportion of green pixels present in an image.

The metrics extraction process, needed to obtain the histograms that generate CPMs, is done within a parallel workflow, which assigns tasks to multiple CPU cores to enhance scalability. In our performance experiments, we observe reasonable speedup in execution time when increasing the number of available threads from 1 to 4. However, further increments did not yield relevant performance improvements because our workflow becomes mainly bottlenecked from disk read speeds.

We combine our results on color representation, reliability, and uncertainty measurement to produce insightful Chronological Percentage Maps. These CPMs summarize the input data reasonably well, allowing the analyst to distinguish phenological phases across multiple years and accounting for occasional anomalies caused by problems during data capture.

As future work, we plan to apply the analysis technique presented in this paper with different vegetation types and varying climatic conditions. We also intend to explore multi-node execution in an HPC cluster to allow for even faster metrics extraction, as well as to investigate other methods to enhance uncertainty measurement, such as providing a statistical test.

Software and Data Availability

We endeavor to make our analysis reproducible. A companion material hosted in a public GitHub repository at `https://github.com/guilhermealles/phenology-analysis-companion/`, contains the source code, datasets, and instructions to reproduce our results. An archive is also available in Zenodo at `https://zenodo.org/record/3771710/`. The code snippets in R (using packages from `tidyverse` and `cowplot`) are capable of reproducing each figure from the paper.

Acknowledgements

We thank the Phenological Eyes Network (PEN) for providing high-resolution datasets (`http://www.pheno-eye.org`). This material is based upon work supported by the FAPERGS Research Foundation (grants GreenCloud 16/488-9 and MultiGPU 16/354-8), the National Council for Scientific and Technological Development (CNPq) (grant 447311/2014-0), the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES) - Finance Code 001, the CAPES/Brafitec EcoSud 182/15, and the CAPES/Cofecub 899/18 projects.

References

Alberton, B., da S. Torres, R., Cancian, L.F., Borges, B.D., Almeida, J., Mariano, G.C., dos Santos, J., Morellato, L.P.C., 2017. Introducing digital cameras to monitor plant phenology in the tropics: applications for conservation. Perspectives in Ecology and Conservation 15, 82 – 90. doi:10.1016/j.pecon.2017.06.004.

Bater, C.W., Coops, N.C., Wulder, M.A., Hilker, T., Nielsen, S.E., McDermid, G., Stenhouse, G.B., 2011. Using digital time-lapse cameras to monitor species-specific understorey and overstorey phenology in support of wildlife habitat assessment. Environmental Monitoring and Assessment 180, 1–13. doi:10.1007/s10661-010-1768-x.

Beck, P.S., Atzberger, C., Høgda, K.A., Johansen, B., Skidmore, A.K., 2006. Improved monitoring of vegetation dynamics at very high latitudes: A new method using modis ndvi. Remote Sensing of Environment 100, 321 – 334. doi:10.1016/j.rse.2005.10.021.

Cheng, H., Jiang, X., Sun, Y., Wang, J., 2001. Color image segmentation: advances and prospects. Pattern Recognition 34, 2259 – 2281. URL: http://www.sciencedirect.com/science/article/pii/S0031320300001497, doi:https://doi.org/10.1016/S0031-3203(00)00149-7.

Chmielewski, F.M., Rötzer, T., 2001. Response of tree phenology to climate change across europe. Agricultural and Forest Meteorology 108, 101 – 112. doi:10.1016/S0168-1923(01)00233-7.

Graham, E.A., Yuen, E.M., Robertson, G.F., Kaiser, W.J., Hamilton, M.P., Rundel, P.W., 2009. Budburst and leaf area expansion measured with a novel mobile camera system and simple color thresholding. Environmental and Experimental Botany 65, 238 – 244. doi:10.1016/j.envexpbot.2008.09.013.

HE, A., S, E., WL, K., R, S., et al., 2009. Tree phenology and carbon dioxide fluxes: use of digital photography for process-based interpretation at the ecosystem scale. Clim Res 39, 261–274. doi:https://doi.org/10.3354/cr00811.

Ide, R., Oguma, H., 2010. Use of digital cameras for phenological observations. Ecological Informatics 5, 339 – 347. doi:10.1016/j.ecoinf.2010.07.002. special Issue on Advances of Ecological Remote Sensing Under Global Change.

Leite, R.A., Schnorr, L.M., Almeida, J., Alberton, B., Morellato, L.P.C., Torres, R.d.S., Comba, J.a.L., 2016. Phenovis - a tool for visual phenological analysis of digital camera images using chronological percentage maps. Inf. Sci. 372, 181–195. doi:10.1016/j.ins.2016.08.052.

Lin, S., Fortuna, J., Kulkarni, C., Stone, M., Heer, J., 2013. Selecting semantically-resonant colors for data visualization. Computer Graphics Forum 32, 401–410. doi:10.1111/cgf.12127.

Morisette, J.T., Richardson, A.D., Knapp, A.K., Fisher, J.I., Graham, E.A., Abatzoglou, J., Wilson, B.E., Breshears, D.D., Henebry, G.M., Hanes, J.M., Liang, L., 2009. Tracking the rhythm of the seasons in the face of global change: phenological research in the 21st century. Frontiers in Ecology and the Environment 7, 253–260. doi:10.1890/070217.

Nagai, S., Akitsu, T., Saitoh, T.M., Busey, R.C., Fukuzawa, K., Honda, Y., Ichie, T., Ide, R., Ikawa, H., Iwasaki, A., Iwao, K., Kajiwara, K., Kang, S., Kim, Y., Khoon, K.L., Kononov, A.V., Kosugi, Y., Maeda, T., Mamiya, W., Matsuoka, M., Maximov, T.C., Menzel, A., Miura, T., Mizunuma, T., Morozumi, T., Motohka, T., Muraoka, H., Nagano, H., Nakai, T., Nakaji, T., Oguma, H., Ohta, T., Ono, K., Pungga, R.A.S., Petrov, R.E., Sakai, R., Schunk, C., Sekikawa, S., Shakhmatov, R., Son, Y., Sugimoto, A., Suzuki, R., Takagi, K., Takanashi, S., Tei, S., Tsuchida, S., Yamamoto, H., Yamasaki, E., Yamashita, M., Yoon, T.K., Yoshida, T., Yoshimura, M., Yoshitake, S., Wilkinson, M., Wingate, L., Nasahara, K.N., 2018. 8 million phenological and sky images from 29 ecosystems from the arctic to the tropics: the phenological eyes network. Ecological Research 33, 1091–1092. doi:10.1007/s11284-018-1633-x.

Pettorelli, N., Vik, J.O., Mysterud, A., Gaillard, J.M., Tucker, C.J., Stenseth, N.C., 2005. Using the satellite-derived ndvi to assess ecological responses to environmental change. Trends in Ecology & Evolution 20, 503 – 510. doi:10.1016/j.tree.2005.05.011.

Pietikainen, M., Nieminen, S., Marszalec, E., Ojala, T., 1996. Accurate color discrimination with classification based on feature distributions, in: Proceedings of 13th International Conference on Pattern Recognition, pp. 833–838 vol.3. doi:10.1109/ICPR.1996.547285.

Richardson, A.D., 2018. Tracking seasonal rhythms of plants in diverse ecosystems with digital camera imagery. New Phytologist 0. doi:10.1111/nph.15591.

Richardson, A.D., Braswell, B.H., Hollinger, D.Y., Jenkins, J.P., Ollinger, S.V., 2009. Near-surface remote sensing of spatial and temporal variation in canopy phenology. Ecological Applications 19, 1417–1428. doi:10.1890/08-2022.1.

Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40, 99–121. doi:10.1023/A:1026543900054.

Sonnentag, O., Hufkens, K., Teshera-Sterne, C., Young, A.M., Friedl, M., Braswell, B.H., Milliman, T., O'Keefe, J., Richardson, A.D., 2012. Digital repeat photography for phenological research in forest ecosystems. Agricultural and Forest Meteorology 152, 159 – 177. doi:10.1016/j.agrformet.2011.09.009.

Walther, G.R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T.J.C., Fromentin, J.M., Hoegh-Guldberg, O., Bairlein, F., 2002. Ecological responses to recent climate change. Nature 416, 389–395. doi:10.1038/416389a.

Yan, D., Zhang, X., Nagai, S., Yu, Y., Akitsu, T., Nasahara, K.N., Ide, R., Maeda, T., 2019. Evaluating land surface phenology from the advanced himawari imager using observations from modis and the phenological eyes network. International Journal of Applied Earth Observation and Geoinformation 79, 71 – 83. doi:10.1016/j.jag.2019.02.011.

Zhang, X., Friedl, M.A., Schaaf, C.B., Strahler, A.H., Hodges, J.C., Gao, F., Reed, B.C., Huete, A., 2003. Monitoring vegetation phenology using modis. Remote Sensing of Environment 84, 471 – 475. doi:10.1016/S0034-4257(02)00135-9.