Information Sciences 372 (2016) 181–195

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

PhenoVis – A tool for visual phenological analysis of digital camera images using chronological percentage maps

Roger A. Leite[a], Lucas Mello Schnorr[a], Jurandy Almeida[b,d], Bruna Alberton[c], Leonor Patricia C. Morellato[c], Ricardo da S. Torres[d], João L.D. Comba[a,*]

[a] Institute of Informatics, Federal University of Rio Grande do Sul – UFRGS, Porto Alegre RS–91501-970 Brazil
[b] Institute of Science and Technology, Federal University of São Paulo – UNIFESP, São José dos Campos, SP–12247-014 Brazil
[c] Dept. of Botany, São Paulo State University – UNESP, Rio Claro, SP–13506-900 Brazil
[d] Institute of Computing, University of Campinas – UNICAMP Campinas, SP–13083-852 Brazil

ARTICLE INFO

Article history:
Received 18 January 2016
Revised 3 August 2016
Accepted 15 August 2016
Available online 16 August 2016

Keywords:
Phenology
Remote sensing
Vegetation index
Visual analytics
Similarity ranking

1. Introduction

Phenology studies the periodic phenomena of plants and their relationship to environmental conditions [39]. This analysis is crucial for accessing the impact on vegetation and ecosystem processes [29,32,39,42]. Examples of cyclic phenomena include flowering and fruiting in plants and the breeding season of birds and frogs, among others. The remote monitoring of vegetations using cameras has proved to be a promising approach to the study of plant phenology [30,38,40]. In this scenario, cameras capture daily pictures from a specific viewpoint at a specific time of the day. By comparing a sequence of images over time, it is possible to identify changes that are associated with phenological events [38]. For example, images that have a high number of pixels with dominant shades of green are often associated with areas mostly covered by leaves.

Due to the large number of images (at least 365 images per year), the visual analysis of large collections of images becomes too complex to be performed interactively. Instead, a single chromatic value is computed to represent the average color in each image. Among several chromatic coefficients described in the literature, the green chromatic coefficient (g_{cc}) is widely used by the phenology community to understand periodic leafing patterns extracted from digital images [40]. The collection of values computed in the period of interest (usually one year) is displayed as a 2D line plot [3], and the analysis

* Corresponding author. Fax: +55 (51) 3308-7308.
E-mail addresses: rogeraleite@gmail.com (R.A. Leite), schnorr@inf.ufrgs.br (L.M. Schnorr), jurandy.almeida@unifesp.br (J. Almeida), bru.alberton@gmail.com (B. Alberton), pmorella@rc.unesp.br (L.P.C. Morellato), rtorres@ic.unicamp.br (R.d.S. Torres), comba@inf.ufrgs.br, joao.comba@gmail.com (J.L.D. Comba).

http://dx.doi.org/10.1016/j.ins.2016.08.052

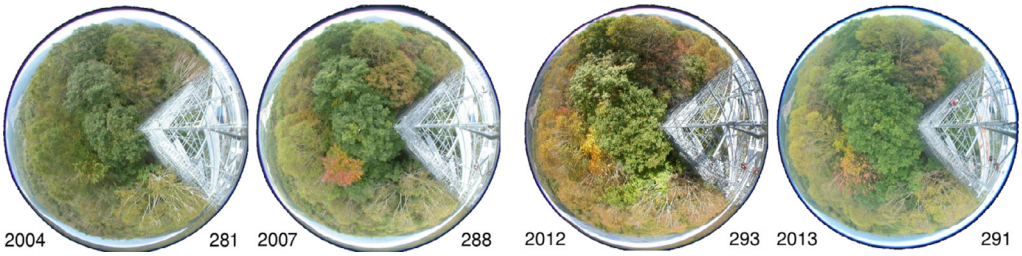


Fig. 1. Images of the Takayama forest from different years (2004, 2007, 2012, and 2013) and days (281, 288, 293, and 291), generated by the Phenological Eyes Network (PEN). Although the images look very different, they have the same average $g_{cc} = 0.3905$.

of this plot is used to identify interesting phenological events. It is well known that using the average as a comparison metric to analyze the images can potentially overlook important patterns and thus lead to wrong conclusions [14]. One motivation for this research is to address the shortcomings of using average scores in near-surface remote phenology.

Fig. 1 illustrates the problem with the average in four different images captured by the same device on different dates. Each image looks different when compared to the others, but all of them share the same g_{cc} average of 0.3905 (the pixels associated with the metallic structure are discarded). Therefore, by using the traditional average-based approach, phenologists will be led to conclude that these four images are the same even if they are completely different in reality. The similar distribution of the shades of green, yellow and brown in the images indicates that 2012 was a full fall season, whereas in 2004 the levels of green were much higher (Fig. 1).

In this work, we introduce PhenoVis, a visual analytics tool that aims at providing insightful ways to analyze phenological data. The main idea behind PhenoVis is the creation of more expressive visual encodings for phenological data visualization. We introduce the Chronological Percentage Maps (CPMs), a visual mapping technique that combines derived distributions from all images of a given year to create normalized stacked bar charts. Another problem we address is the lack of automation in the search for regular patterns in plant phenology. Current research in the field frequently uses the analyst's intuition to find patterns of plant phenology. This approach is potentially time-consuming, error-prone and unable to scale to larger datasets. PhenoVis uses the additional information encoded to support similarity searches, which is useful for comparing data from different years. It also provides a customizable multi-rank comparison of data from different years, with filters of specific periods within a year or sub-regions associated with given plant species.

In summary, we offer the following contributions: (a) PhenoVis, a visual analytics tool to perform a comparative analysis of phenological data for multiple years; (b) the design of CPMs, in which a more expressive representation of phenological data using percentage distributions is combined with a visual expression of this information using color-coded normalized stacked bar charts; (c) similarity algorithms that allow the search for similar phenological patterns across years, occurring in either a fixed or moving window of time; (d) customizable ranking comparison of years, with filters that allow selections of specific time periods or regions; and (e) case studies that validate PhenoVis in phenological analysis tasks such as pattern and outlier identification.

2. Related work

Below, we review visual analytics and information visualization techniques and approaches to analyse phenological data extracted from images.

2.1. Visualization and data Analysis

There is an extensive literature in the information visualization and visual analytics community on different ways to visually present information and to extract patterns from data [25]. Rectangular regions such as matrices are a common alternative for presenting aggregated data to allow identification of patterns. Pixel bar charts extend traditional bar charts by coloring pixels inside bars to represent information derived from data attributes [26]. The visual presentation of data follows a pixel-placement ordering, which was demonstrated to reveal patterns in data [24] as well as provide support for data mining queries. The underlying motivation behind pixel bars was an inspiration when designing PhenoVis, particularly the CPM representation and the ability to search for similar patterns in CPMs of different years.

There are several works that use matrix-encoded information to perform visual analysis. Stacked bar charts are used to display the temporal changes of traffic speed data in [8]. The normalized stacked bar charts used in CPMs resemble the images showed in their work, especially when using a categorical color mapping. The analysis of visual traffic also appears in [43], with trajectory information chronologically encoded into a matrix. Matrices are used to compare genomic sequences in [2,34]. BallotMaps encode the votes received by politicians into matrices to identify voting patterns [46]. The Flowstrates approach encodes origin-destination data into a heatmap matrix, re-ordering rows to reveal interesting patterns [9]. In [36], a matrix encoding heart-rate information is used to identify unusual patterns during a running race. The exploration of dynamic graphs using matrix-encoded information is given in [12] and [18]. Climate change comparison using a global

radial map is presented in [27]. Although it uses a radial representation, the compact representation resembles the CPM stacked bar charts. ThemeRiver [11,19,20] offers an alternative to stacked bar charts by displaying information in layers that are stacked in a symmetrical shape centered around the x-axis. This approach allows easy identification of predominant layers in time, particularly for datasets comprising a great number of layers. LineUp describes different ways to present multi-ranking attributes [17]. The multi-ranking visualization used in our work reveals the need for ranking visualization when performing comparative phenology analysis. We refer the reader to [1] for a comprehensive survey of other ways to perform visual analysis of time-oriented data.

2.2. Phenological analysis

Phenology analysis of satellite images often relies on average plots of computed vegetation indices. A web-based interface described in [10] displays time-series obtained from phenological and meteorological observations. TimeStats is a free software that offers tools for the visualization of long-term remote sensing data archives such as parametric and non-parametric methods for trend detection, linear regression, and frequency analysis [41]. EcoIP, a toolkit to estimate the onset and ending dates of phenological phases of plant species, relies on a Naïve Bayesian model created from a set of training images, which is used to provide temporal estimators [16]. A variety of color transformations is used to adjust the accuracy of the estimations. Another recent trend refers to the construction of toolboxes. Eerens, for example, developed SPIRITS, a stand-alone toolbox to produce evidence-based information for crop production analysts [13]. It includes a large number of features to analyze image time series and to create maps and graphs for vegetation status analysis. Different from those initiatives, our proposal relies on visual encodings designed to capture distribution aspects from vegetation data. A similar approach is taken in [5] and [6], which presented different strategies for encoding image time series as visual rhythms [33]. Such representations have proven to be a powerful tool for distinguishing the behavior of different plant species. The visual rhythm construction is similar to the CPM extraction process in the sense that they summarize image sequences in a single image representation. Despite the good results observed concerning the use of visual rhythm, no visual analytics tool has been proposed.

3. PhenoVis

PhenoVis is a visualization tool that includes a visual mapping representation called CPM, data analysis algorithms for identifying similar patterns in different years, and a ranking visualization module to present the similarity results. In this section, we describe each of these modules.

3.1. Chronological percentage maps (CPMs)

CPM is the main concept behind PhenoVis. Its construction consists of six steps that transform a sequence of images into a normalized stacked bar chart, as shown in Fig. 2. We detail these steps in the following sections.

Step 1: Filtering by region of interest (ROI)

The region of interest (ROI) (also called mask) is a portion of the input image that is the focus of the analysis, thus removing irrelevant areas that lack vegetation such as the observation tower. The ROI is user-configurable and implemented through a mask with the same dimensions as the input images. The mask is composed of black and white pixels (a stencil image), where white represents selected regions. Two types of masks are used in PhenoVis: a community mask [21], which considers all plant species in the image; and a species mask [30], associated with a given plant species.

Step 2: Data transformation using a vegetation index

The phenophases and length of the growing season (from start to end) are indicators of plant development across different years. The phenological analysis from images focuses on the following four checkpoints (or phenophases) of plant development [21,30]:

- Leaf expansion: plants start to turn green, and leaves are expanding;
- Peak or maturity: leaf growth reaches a plateau; i.e., leaves reach full size;
- Leaf fall or senescence: leaves change colors and start to fall;
- Post leaf fall: there are no leaves on the tree, representing the end of the growing season.

Among the several images taken at different times of the day, the image taken at noon is preferred for the analysis because it minimizes shadow effects. Therefore, only one image per day is used. The analysis considers the chromatic coefficients associated with each pixel in the image. Different vegetation indexes are described in [40]. The *RGB* chromatic coefficients (r_{cc}, g_{cc} and b_{cc}) are defined by respectively dividing each component (R, G, or B) by the sum of the other components ($R+G+B$). The average plot of g_{cc} is a good indicator of phenophases due to the encoding of green pigments (e.g., chlorophyll) in leaves. Although most of our analysis used the g_{cc} chromatic coefficient, we also experimented with the *hue* index discussed in [15]. To use this index, it is necessary to convert colors from the RGB color space into the HSV (also called HSI) color space [44], which represents colors as hue (H), saturation (or lightness) (S), and value (or brightness) (V). HSV is

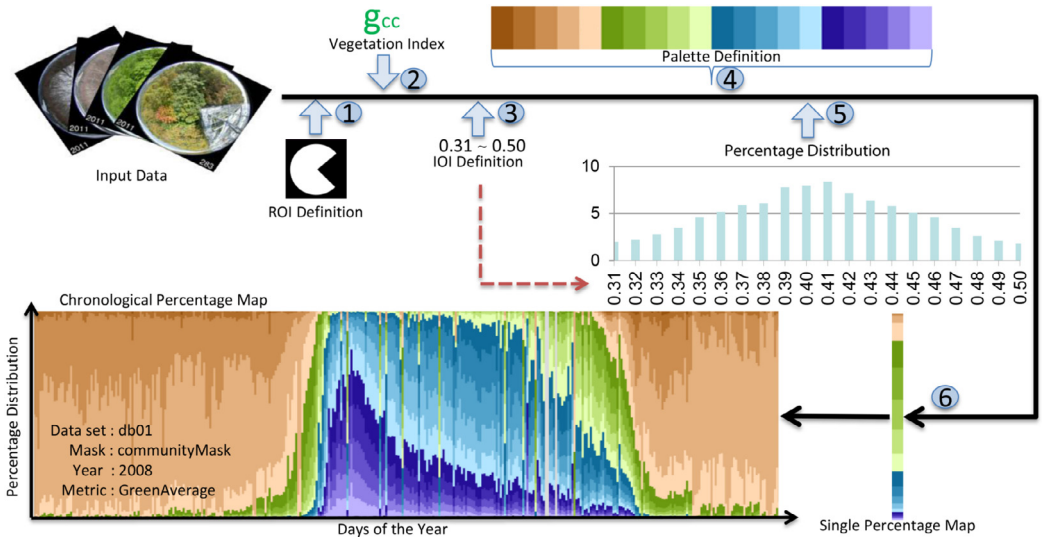


Fig. 2. The Chronological Percentage Map's six-step construction: (1) images are filtered by the community mask that defines the region of interest (ROI); (2) phenology metric (g_{cc}) is computed; (3) interval of interest (IOI) filters the resulting values; (4) selection of the distribution granularity and associated color values; (5) percentage distribution is computed; (6) mapping color values to the percentage distribution. Each image generates a stacked bar chart associated with a single line of the CPM. In this example, all lines of the CPM are stacked in landscape mode.

defined in a cone using cylindrical coordinates, with shading variations defined by the angle around the central vertical axis of the cone (the hue component).

Step 3: Filtering by interval of interest (IOI)

Each vegetation index has values in a particular range. For example, g_{cc} is normalized between 0 and 1. In practice, the range of g_{cc} found in vegetation images is much smaller, usually between 0.3 and 0.5 [47]. Since we compute a histogram of vegetation indexes, we narrow the limits of the histogram to a user-specified interval of interest (IOI), as shown in Fig. 2, step 3.

Step 4: Color palette and histogram granularity

Selecting a good color palette is essential in data analysis [28]. In PhenoVis we associate different colors with each bucket of the percentage histogram. The histogram granularity defines the size of a given bucket of the percentage distribution. The number of buckets is given by the number of colors available, and the range of the distribution is given by the IOI. There are trade-offs when selecting the size of the IOI, granularity, and the number of colors. For example, if the interval is small and the number of colors is large, each dimension of the IOI will be semantically irrelevant because of the resultant tiny grain size. The same problem appears if the interval is too large and the number of colors is small. The best situation arises when there is a balance between the size of the IOI and the number of colors. The typical IOI for the g_{cc} varies from 0.3 to 0.5. Therefore, we chose 20 colors, with each bucket size being 0.01 wide.

For convenience, we have pre-defined a set of palettes. From this set, the categorical color table in Fig. 2, step 4 was the preferred choice when using g_{cc} indexes. The palette has four categories with different colors. Each of the five internal divisions of these four zones has distinct levels of saturation (from dark to light brown, for example). On the other hand, the standard hue color table is the choice when using *hue* indexes.

Step 5: Calculating the percentage data distribution

Each pixel in a given input image contributes one vegetation index value to the analysis, and when this value is multiplied by the number of images in a year (365), a huge amount of data is generated. Instead, the average of all indexes in an image is used, leading to only 365 indexes. Average plots of these indexes help identify phenophases; however, this analysis may be misleading since different images might have the same index (see Fig. 1). In PhenoVis, we store more information about each image than a single vegetation index. We use a histogram instead to encode the percentage distribution of vegetation indexes in an image. The range of the histogram is defined by the IOI, and the number of buckets by the size of color palette used. For indexes outside the IOI, we either discard the values or clamp them to the nearest extreme bucket. We repeat this process to compute one percentage distribution for each image.

In some situations two distinct RGB colors may have the same g_{cc} (Fig. 3), and would therefore be accumulated in the same histogram bin and mapped to the same color. In these situations, the *hue* index can be used alternatively. The corresponding hue values in Fig. 3 are different and can be mapped to distinct colors. Shades of green can be found around

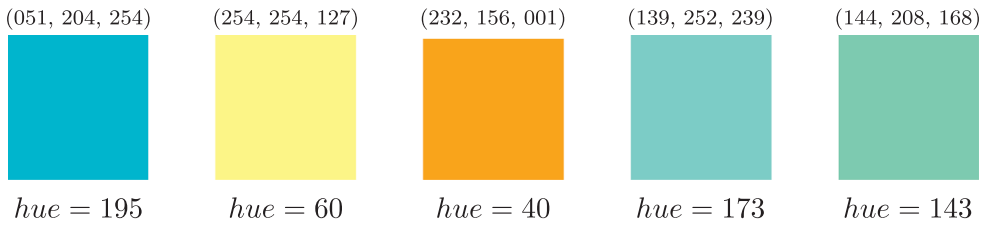


Fig. 3. Colors with the same g_{cc} of 0.4 (RGB codes are shown above each color). On the other hand, the corresponding *hue* values are distinct. By using the *hue* values to separate these colors, the histogram can better aggregate shades of green into closer bins.

hue values of 120. In addition, the mapping of hues into colors can use the standard HSV palette at fixed S and V values. We did not replace g_{cc} by the *hue* index, but used it to gain additional insights when looking at the data.

Step 6: Creating the normalized stacked bar chart

The percentage map of a single input image consists of a normalized stacked bar chart (vertical bar in Fig. 2, step 6). In this example, the height of the chart is proportional to the frequency count in the percentage distribution of the vegetation indexes. The width is associated with the number of pixels available for drawing the percentage map. Colors are defined by the palette being used and their corresponding histogram entries.

The chronological percentage map consists of a sequence of percentage maps stacked in chronological order, from top to bottom (portrait) or left to right (landscape). In our analysis, each CPM corresponds to all images from a single year. Fig. 4 and Fig. 5 demonstrate the percentage map and CPM chart for images captured in Japan (the TKY dataset [31,35]). Fig. 4 illustrates percentage maps for individual days of 2006; each map uses two vegetation indexes. The first one uses g_{cc} and a categorical color table where shades of green, blue, and purple respectively correspond to g_{cc} values between 35 – 40%, 40 – 45%, and 45 – 50%. The second one uses the *hue* index and the hue color mapping. The color mapping is applied over the input camera to illustrate the shading variations for each chromatic coefficient. Fig. 5 shows the complete CPM for all days in 2006. Higher g_{cc} values are associated with greener regions, and a clear pattern of growing season emerges in the middle of the year. In the hue mapping, the growing season is associated with shades of green. Fig. 5 illustrates CPM charts that use different vegetation indexes for the year 2006. This CPM allows us to investigate interesting patterns outside the growing season, which displays shades of red and blue.

3.2. Multi-year phenological analysis

The comparative analysis of data from different years is one question that drives plant phenological analysis. The comparison of the starting dates of a given phenophase (e.g. start of leaf growing season) allows us to identify how these values changed with the progression of the years. Automated similarity comparison of different phenophases is important in this process. PhenoVis provides an automatic similarity analysis, because more data is encoded in percentage distributions. The user can select a time interval directly over the CPM of one year, and PhenoVis will suggest a similarity rank of the other years based on this pattern. In this section, we describe how PhenoVis performs the similarity search and displays the visualization results.

3.2.1. Search period and ROIs

PhenoVis allows the user to define any period of a given year to serve as the basis of comparison against other years. This period of interest used for searches can be manually configured in the interface by specifying start and end dates. Since most queries are related to the analysis of pattern variations in specific phenophases, such as the leaf expansion or fall period, we allowed the user to select the period of interest from pre-defined phenophases. The search can be performed over the ROI of all plant specifies in the image (community mask), other pre-defined masks of any given species, or any region in general specified by the user.

3.2.2. Similarity metrics

PhenoVis offers four similarity techniques for automatic search: Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), and the Kullback-Leibler Divergence (KLD) as described in the supplementary material. The MAE is a usual estimator of the difference between two matrices. Used for the same reasons, MSE highlights significant gaps between values due to squared distance computations and is not recommended for noisy datasets. MAPE is a normalized solution based on the percentage of the similarity for two samples, while the KLD is a non-symmetric measure of the difference between probability distributions. The similarity is computed using two CPMs as input. The first parameter is the CPM query, defined by its start and end dates. The second parameter is the CPM candidate, which will necessarily have the same number of days as the target, but the start date may differ. The percentage distribution associated with each CPM subset is interpreted as a matrix of values in which the similarity metric is computed. Errors are computed for each matrix entry and accumulated in the resulting similarity error.

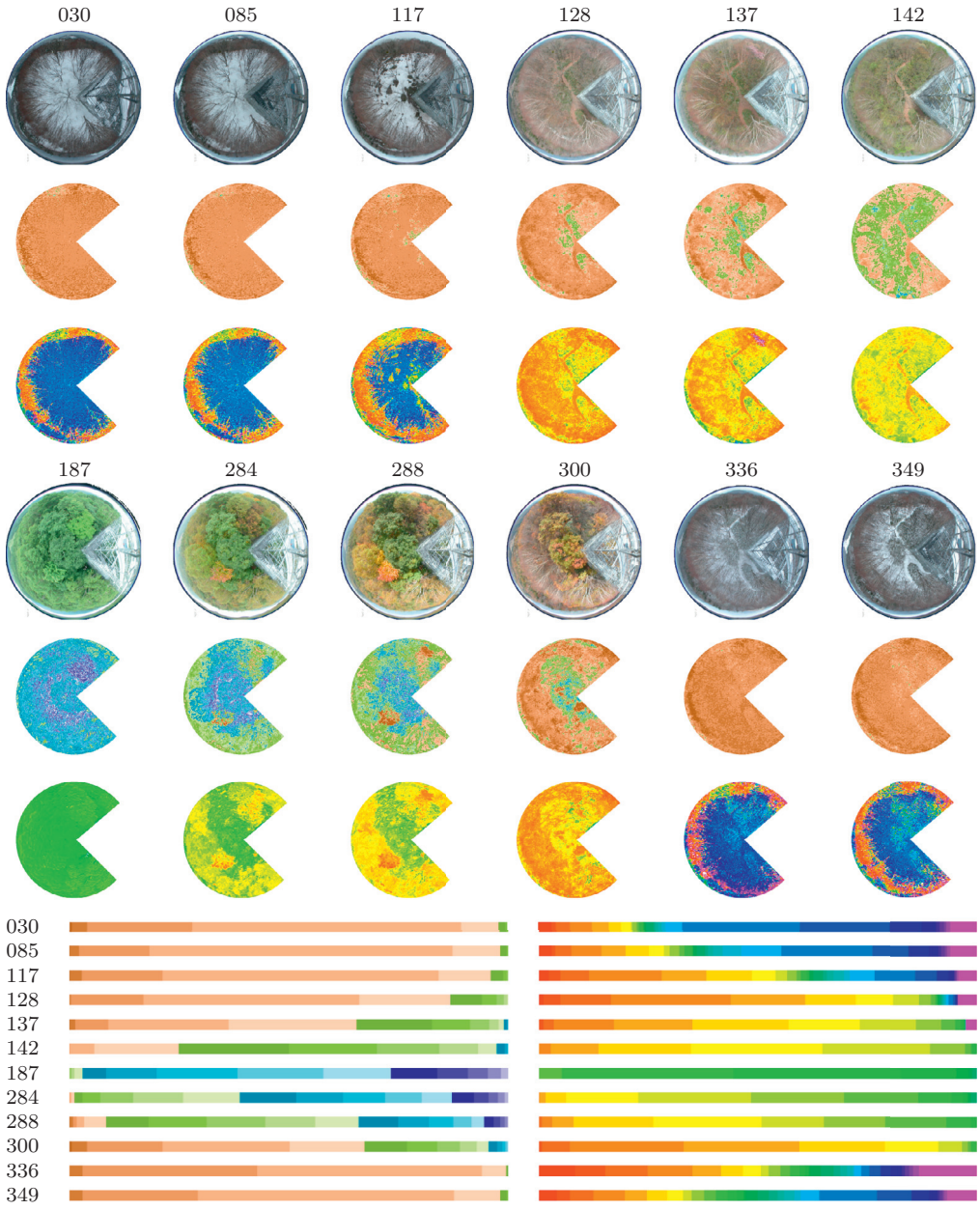
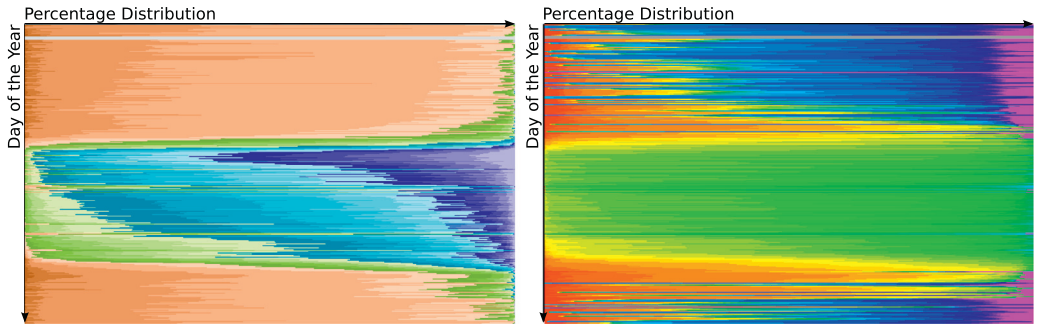


Fig. 4. Percentage maps for 12 days: original data for the respective days (row 1 and 4), recoloring of pixels using g_{cc} and a brown-green-blue-purple color mapping (rows 2 and 5), and recoloring using the hue index and hue mapping (rows 3 and 6). The percentage maps for the g_{cc} (left) and hue (right) indexes are shown.

3.2.3. Searching windows and filters

The query defines a time period in the CPM of a given year. The search for similar patterns looks for CPM subsets that have the same number of days as the query pattern. This search is implemented in two configurable ways. The first one, referred to as fixed window, looks for CPM subsets with matching start and end dates. The result of this comparison gives the years in which the same pattern occurred on the same days of the year. In the second way, referred to as moving window, the search looks for the same pattern but does not fix the starting day, allowing the window to move throughout the year. This search finds patterns that happened at a different time of the year (e.g., a late growing season) in other years.

We also implemented a filter that allows a subset of the percentage distribution to be considered. If the analysis is interested only in g_{cc} values in the interval of 40% to 45%, the similarity search can be set to filter out values outside this interval. We also have an outlier filter that removes outliers (missing data or odd days) from the similarity search. An



(a) g_{cc} (b) hue

Fig. 5. CPM example for 365 days displayed in portrait mode. Each line corresponds to a normalized stacked bar chart mapped to the predefined color table: (a) categorical color table using the g_{cc} index, (b) HSV color table using the *hue* index.

automatic outlier detection was implemented by taking into account the percentage distribution of a given day and the distribution of adjacent days.

3.3. Single and multi-year ranking

PhenoVis is also capable of creating a single or multi-year ranking according to the query and selected similarity technique. The ranking includes the distributions for different years in order of similarity and the similarity error, which allows the evaluation of how the years differ among them. The single-year ranking generates a histogram plot that allows the comparison of one specific year's distribution against all others. The goal of this ranking is to compare the phenophases of one year against the others. In the case of a moving window, the ranking also includes the distance (in days) of the closest pattern.

The multi-year ranking creates similarity searches between a pair of ROIs in a given period of time for every pair of years. This process first results several histogram plots, one for each year, that indicate how close or distant the other years are. These histograms can be normalized as needed; in PhenoVis we use three normalizations: single-year normalization, multi-year normalization, and global normalization (using the data for all years and all pairwise comparisons). The resulting histograms are drawn in red, green and blue, respectively. The second result is a histogram plot that summarizes the individual yearly histogram plots by associating a distance from one year to the others. In other words, this histogram displays the most common year and the most different years as distances between each other, which allows us to easily identify the average years. The resultant histogram is drawn in magenta. Fig. 6 illustrates the interface of PhenoVis, and samples of these histograms are shown in Fig. 6(b).

4. Visualization results and discussion

PhenoVis was implemented using the programming language *Processing* [37]. The input to the system is a collection of raw images grouped by year and a mask image that defines the region of interest. In off-line computation, the mask is applied for all input images, and the percentage distributions are computed and stored. PhenoVis was tested, under permission, with the Takayama Flux Site (TKY) dataset of the Phenological Eyes Network (PEN) [31,35][1]. TKY is located in Japan, which has images from multiple years (from 2004 onward) of a deciduous broadleaf forest [30]. For the results below we used the 600 × 600 image resolution database due to its reduced pre-processing time. Experiments conducted with the larger resolution dataset (2272 × 1704) showed similar results to the smaller dataset. We aim to extend our prototype for the larger resolution datasets, possibly using parallel computation to speed-up pre-processing time. We used the same regions of interest defined by [30] to identify three species: *Betula Ermanii*, *Quercus Crispula*, and *Acer Rufinerve*. Fig. 7 illustrates the location of the species in the image. A discussion of the results obtained using the TKY dataset can also be found in [22,23]. Our results use the following phenological phases described in [30] to evaluate the TKY dataset:

- Bud dormancy (BD): from January to early April (days 1 to 100);
- Leaf Expansion (LE): from late May to late June (days 140 to 180);
- Peak period (Peak): from early July to mid-September (days 181 to 258);
- Leaf Fall (LF): from late September to early November (days 263 to 315);
- Post Leaf Fall (PLF): from mid-November to late December (days 319 to 365);

[1] (http://www.pheno-eye.org).

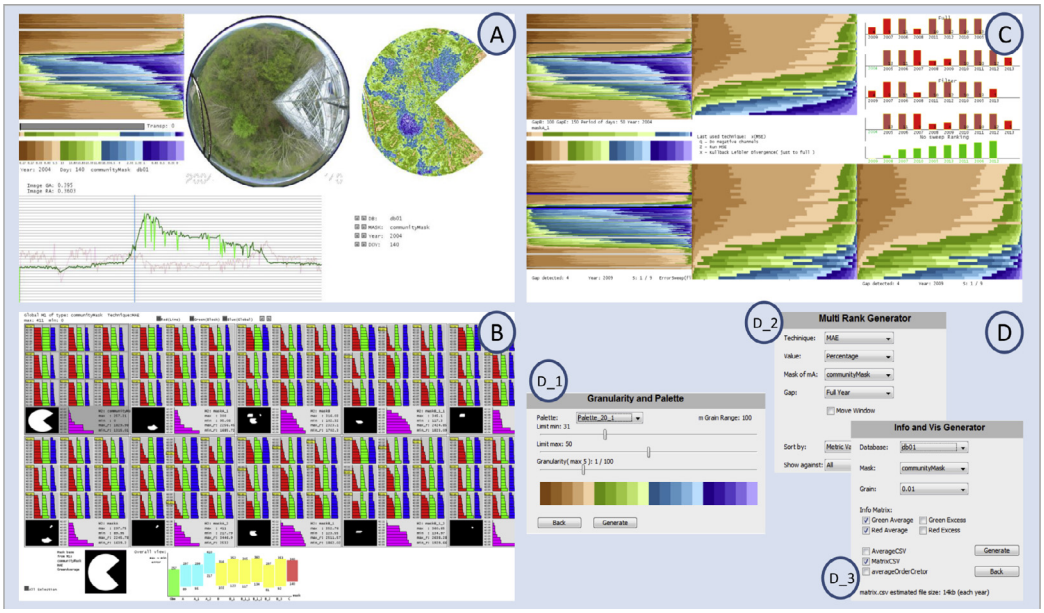


Fig. 6. Selected screenshots of the PhenoVis interface: (A) CPM analysis mode, with interaction to inspect individual images and specify query windows; (B) multi-rank results with pairwise comparison of all years; (C) single-rank results using fixed and moving windows. In (D), we display additional windows for configuring search parameters.

(a) species (b) community (c) A_1 (d) A_2 (e) B_1 (f) B_2 (g) C

Fig. 7. Species location (a) and their masks: (b) community, (c) and (d) *Betula ermanii* (A1 and A2), (e) and (f) *Quercus crispula* (B1 and B2), and (g) *Acer rufinerve* (C)

The realization and evaluation of experiments received feedback from the phenology experts involved in this research. Selected screenshots of the interface are shown in Fig. 6. To increase the reproducibility of our work, we prepared a git repository with the core programs of PhenoVis[2].

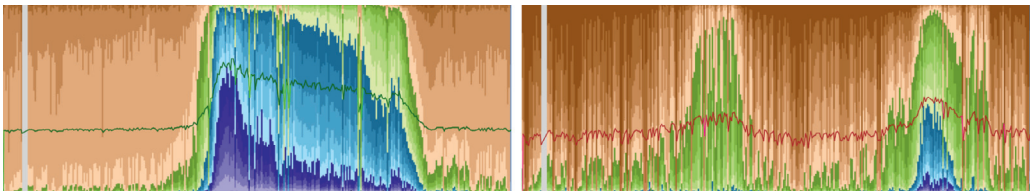
4.1. CPM evaluation

In this section, we compare the CPM representation against average plots used in phenological studies, an evaluation of the CPM expressive power, and a novel outcome that consists of using the CPM as a species signature.

4.1.1. Comparison against average-based plots

We use an example to illustrate how the CPM better displays changes in the data throughout the year than average plots. Fig. 8 shows the average plot based on the g_{cc} and r_{cc} (displayed as a green or red line) over the CPM for 2006. While the average has changed over time, it fails to illustrate the data's composition as CPM does. In the r_{cc} plot, we observe that the shading variations during the leaf-fall period can be easily spotted in the CPM. The red channel indicates the leaf senescence, which occurs in the fall; an expanded discussion on this can be found in [40].

Fig. 9 shows the difference between percentage maps and plots that have the same average. It shows four percentage maps of different days with the same $g_{cc} = 0.3905$. Although they have the same g_{cc} average, the percentage map is clearly different. Percentage maps and CPMs allow us to inspect differences in ways that average plots can not.



(a) g_{cc} average plot and CPM (b) r_{cc} average plot and CPM

Fig. 8. CPM (landscape mode) and the corresponding g_{cc} and r_{cc} for the year (DOY) of 2006. The CPM better expresses the changes in the data than the average plots,

Fig. 9. Four percentage maps with the same g_{cc} average (0.3905). In the CPM representation, they show different aspects. The color palette is described in Step 4, Section 3.1.

4.1.2. CPM expressive power

The side-by-side comparison of CPMs can reveal relevant patterns on the data. In Fig. 10, we give CPMs for the years 2004-2012. We observe that the years of 2004 and 2009 are different from the others. One possible explanation is that 2004 had the highest temperature and humidity indexes of all years. Another possible explanation is that the TKY site was attacked by typhoons. On the other hand, 2009 had the lowest humidity of all years and the lowest snow index, which reveals more of the terrain around trees during winter time.[3] Note that the *hue* index natural color associations have a direct relation to leaf exchange patterns.

4.1.3. CPM as species visual signature

In this section, we evaluate how the expressive power of CPMs can be used as a species' visual signature. For this analysis, we used the manual species identification of [30] for the TKY dataset described before.

The placement of adjacent percentage maps in the CPM generates visual cues along the time axis. By exploiting this characteristic, we can see that different species create different CPM patterns. Fig. 11 shows the CPM for the *Betula*, *Quercus* and *Acer* species using the g_{cc} index for 2007 and 2008. In this analysis, we use g_{cc} and the categorical color table to identify different patterns. The three most relevant g_{cc} intervals in the categorical mapping are $35-40\%$ (colored in green), $40-45\%$ (colored in blue), and $45-50\%$ (colored in purple). The patterns we observe for each species is consistent for the two years. However, they have distinct patterns when comparing one species against the other, especially during leaf growth and senescence phases. *Quercus* has a purple zone that shows a peak for the greening phase, which stabilizes in the rest of the growing season.

4.2. Multi-year data analysis

We investigate in this section the ability of PhenoVis to perform data analysis and search for similar phenological patterns. The identification of inter-annual variability in phenological data series is of key importance to identify changes and trends that can be related to environmental drivers.

4.2.1. Searching for similar phenological patterns

PhenoVis allows the user to specify a window of time over the CPM to define an interesting pattern to be searched. This pattern will be used as a query for similar patterns in other years. Fig. 12 shows the results of a search that uses as its query window the leaf expansion period of 2009 and the MSE similarity metric. The obtained results show that this pattern was most similar to the years 2008 and 2004 when using a fixed window. On the other hand, it was closer to 2007 and 2006 when using a moving window. As we can see, the ranking results can be different when the fixed and moving window searches are compared against each other. Moreover, as expected, the moving window approach presents a smaller error than the fixed window.

[2] https://github.com/schnorr/phenology.

[3] Both temperature and humidity information were obtained at http://www.data.jma.go.jp, using the station TAKAYAMA WMO Station ID:47617 (As of May 2015).

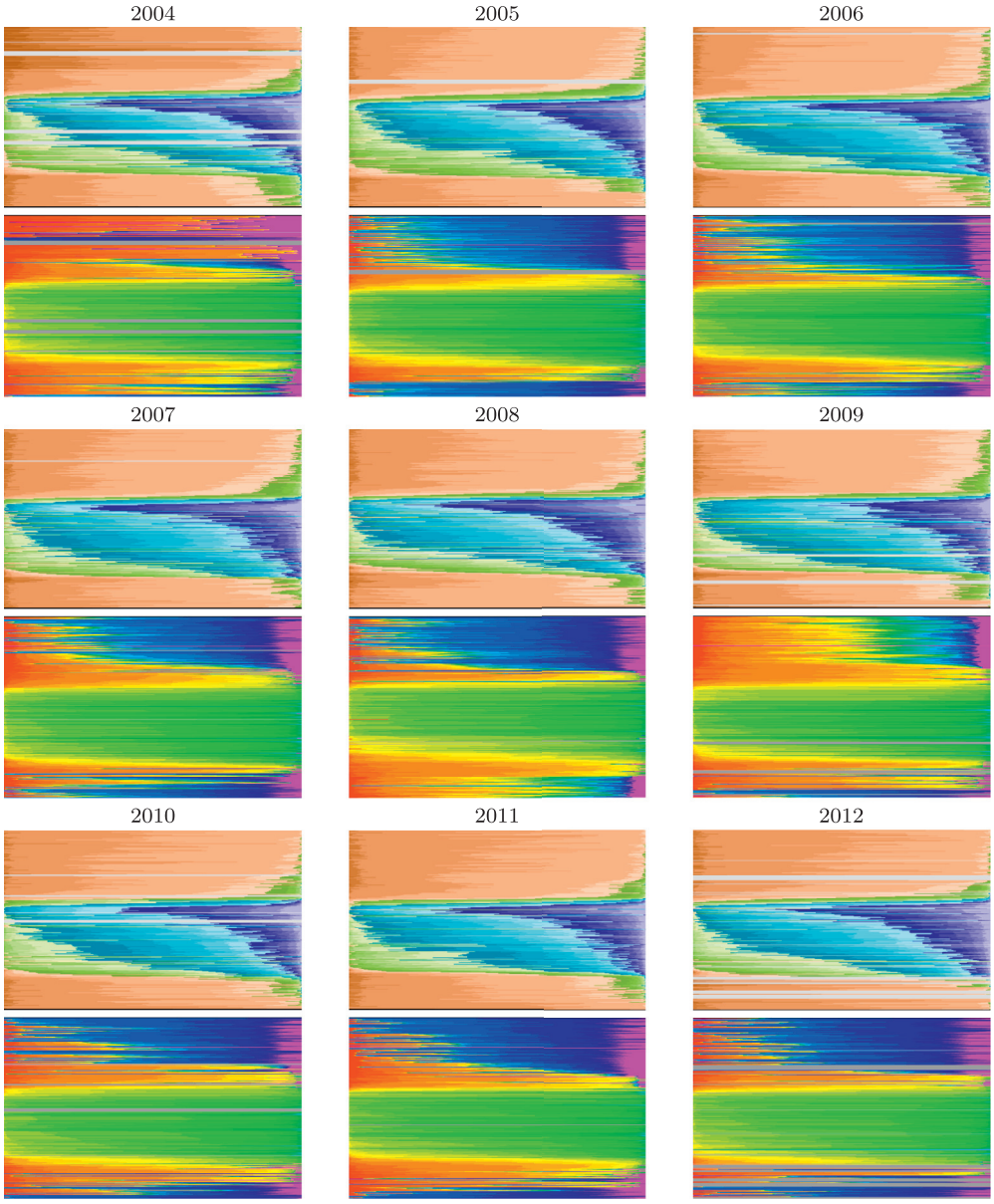
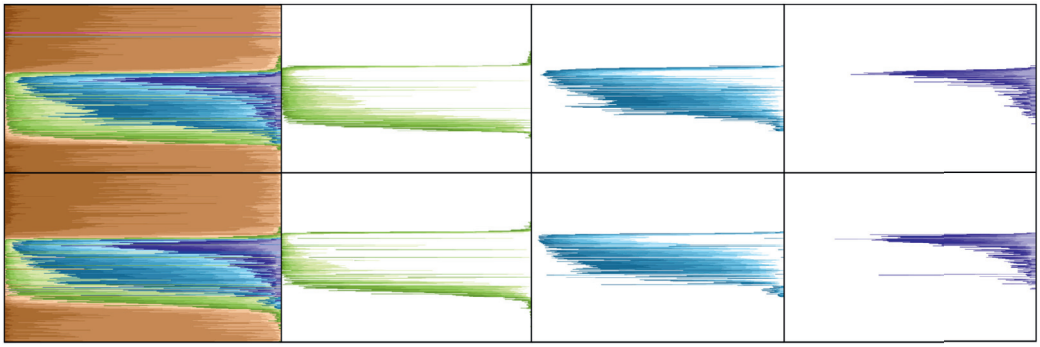


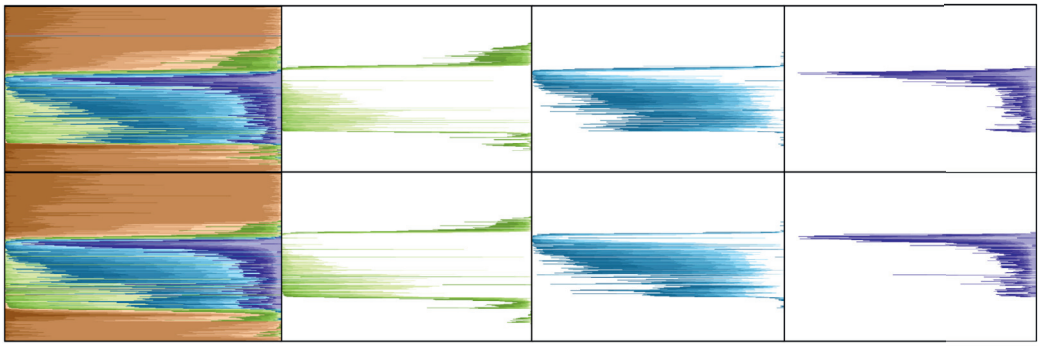
Fig. 10. CPMs for years 2004-2012 using the g_{cc} (rows 1, 3 and 5) and *hue* (rows 2, 4, and 6) indexes. We observe in the hue CPMs that years 2004 and 2009 were distinct from other years, while green colors consistently identify the leafing period.

4.2.2. Multi-year ranking

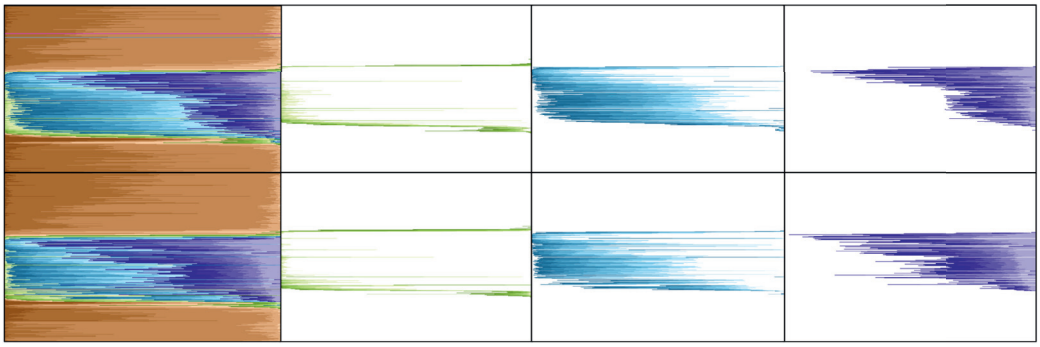
In a single-rank analysis, the user selects a query pattern of a given year. Results are ordered by the distance computed using the similarity metric. The multi-rank performs pairwise comparisons of all years. Fig. 13 shows the results obtained with the multi-year ranking using the entire year as the time interval, and pairwise comparisons of the community and species ROIs. For each ROI (community, *Betula*, *Quercus* and *Acer*), the 9 histograms display in green the normalized distance of every year from 2004 to 2012 against the others. The histogram is sorted according to the error, with the most similar year on top and the most different year on the bottom. For example, in the community mask, the most different years were 2014 (for 5 other years), 2009 (for 3 years) and 2012 (for 1 year). For *Betula*, the most different years were 2012 (for 5 other years), 2009 (for 3 years) and 2004 (for 1 year); for *Quercus*, 2004 (for 8 other years) and 2009 (for 1 year); and for *Acer*, 2004 (for 6 other years) and 2009 (for 3 years). The purple histogram shows the combined information of all the histogram plots of a given ROI in a single summary plot. This summary plot confirms the information seen in the individual plots, which show that 2004, 2009 and 2012 were the most different years, while 2007 was the most similar year. For *Betula*, we observe that the most different year was 2012, as noted previously.



(a) *Betula ermanii*



(b) *Quercus crispula*



(c) *Acer rufinerve*

Fig. 11. CPMs and zone highlights of different species for the years 2007 (top row), and 2008 (bottom). Observe the distinct patterns in the three zones (green, blue, and purple), which can serve as a visual signature to identify a given species.

In Fig. 14, we use the summary plots to refine our analysis to the pre-defined phenophases of the *Betula*, *Quercus*, and *Acer* species. For each phenophase described at the start of this section (BD, LE, Peak, LF and PLF), we display one summary plot. For *Betula*, the most different years in BD, LE, Peak, LF and PLF phenophases were 2004, 2008, 2004, 2012 and again 2012. For *Quercus*, the most different years in BD, LE, Peak, LF and PLF phenophases, were 2004, 2008, 2009, 2009 and 2012, while for *Acer*, they were 2004, 2004, 2009, 2012 and 2008. We also observe that, for BD, the year 2004 was the most different for the three species; for LE, it was 2008 (for two species); for Peak, it was 2009 (for two species); for LF, it was 2012 (for two species); and for PLF, it was also 2012 (for two species). Such analysis is easy to produce with the same encoding and reveals interesting aspects about the data that would be hard to extract from average plots.

In Fig. 15, we used the moving window to search for the specific pattern associated with the leaf expansion period of the years 2007 and 2008. The bar height indicates the number of days that the same pattern appears before or after the query pattern. For example, for 2007, the most similar pattern occurred five days earlier in 2004, two days later in 2005, one day later in 2006, and so forth. With respect to 2007, we observe that the leaf expansion happened earlier in 2004, 2008 and

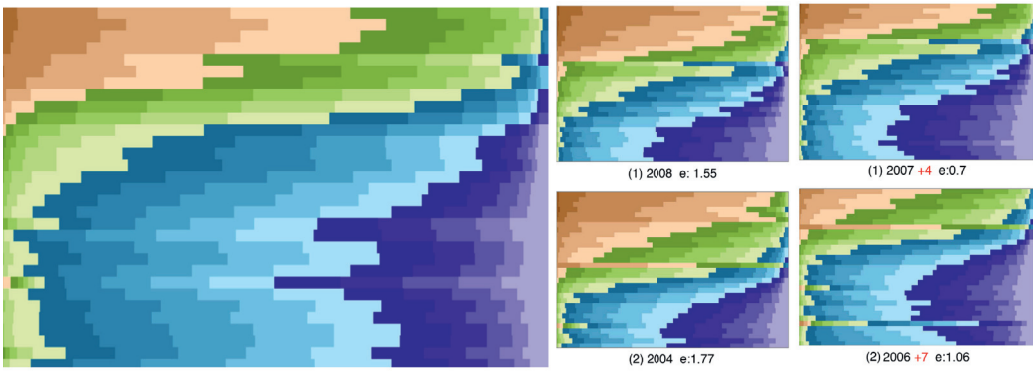


Fig. 12. Searching for similar phenological pattern using MSE: (a) query pattern; and top two results using (b) fixed and (c) moving window. Comparison errors are given after the letter 'e'; red numbers show the temporal shift in days from the matching pattern.

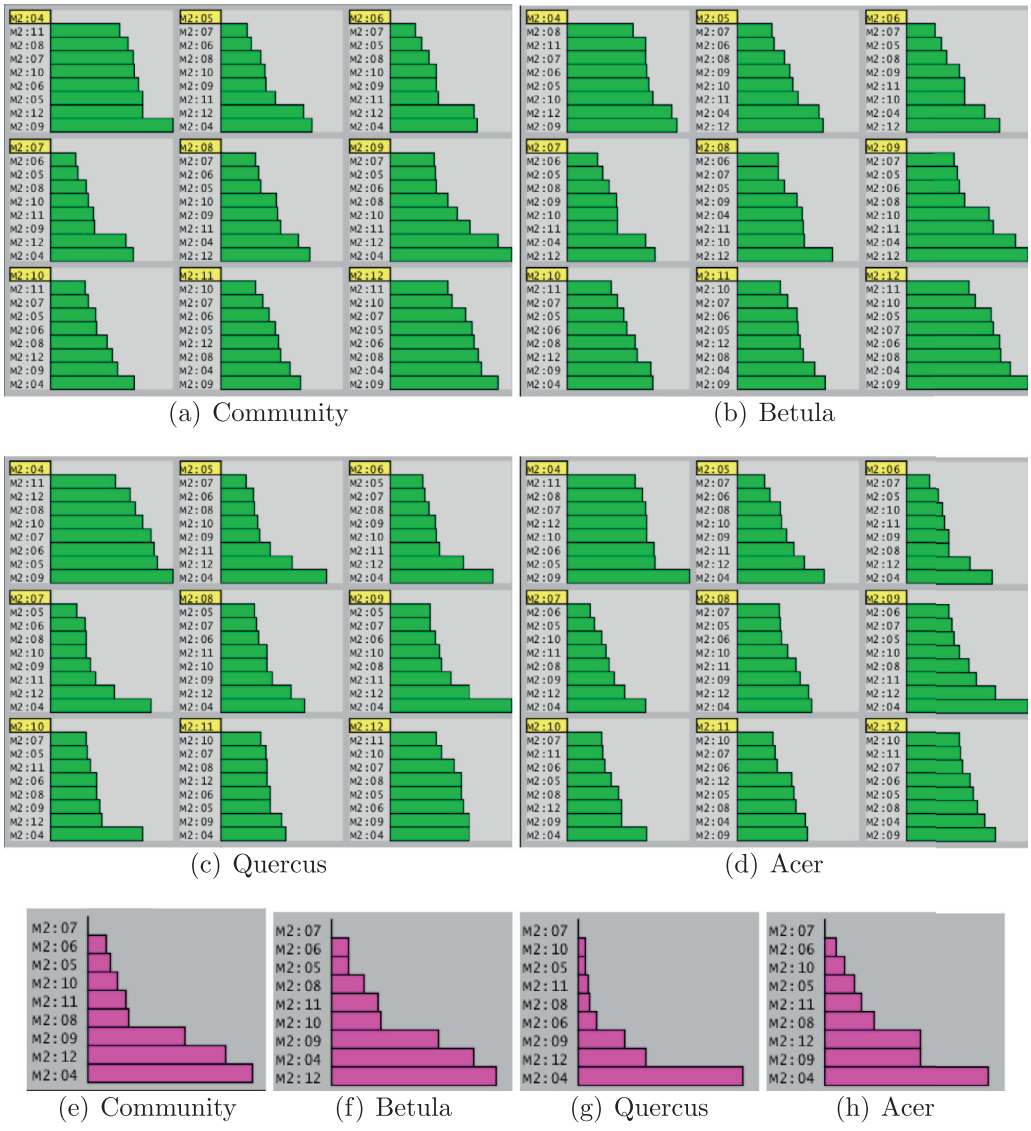


Fig. 13. Multi-rank comparisons of four ROIs (*Community*, *Betula*, *Quercus* and *Acer*) for the entire year: (a)-(d) normalized distance of every year from 2004 to 2012 against the others; (e)-(h) combined information of the green histogram plots of a given ROI into a single summary plot; the year 2007 was the year most similar to the others, while the years 2004, 2012, and 2009 were the most different ones.

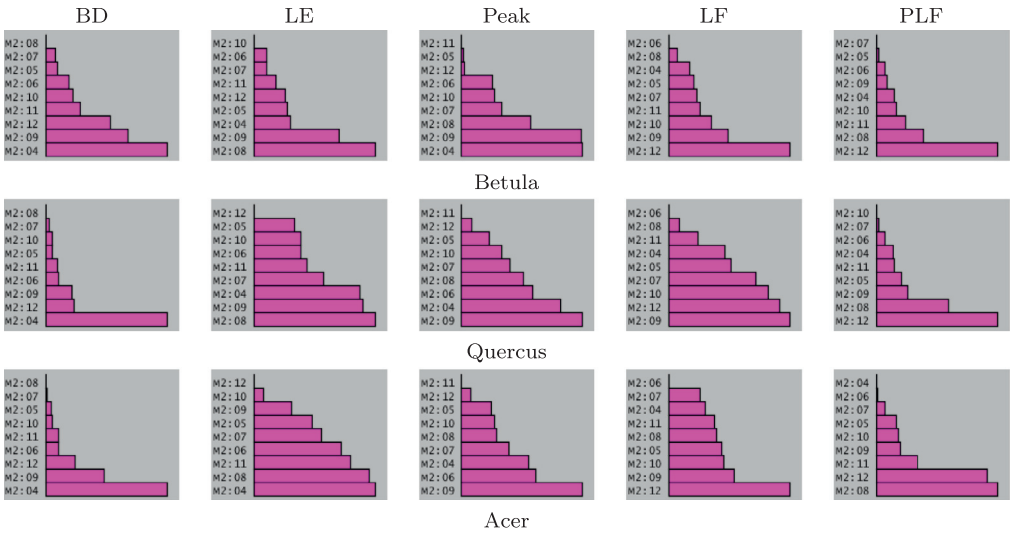


Fig. 14. Summary histograms of the community mask for specific phenological phases. The year 2004 was the most different for the three species; for LE, it was 2008 (for two species); for Peak, it was 2009 (for two species); for LF, it was 2012 (for two species); and for PLF, it was also 2012 (for two species).

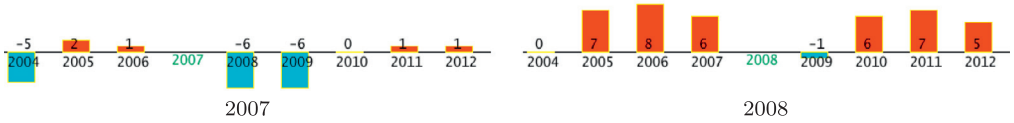


Fig. 15. The leaf-expansion period was used as input pattern in the moving window approach for the years 2007 and 2008: the bar heights represent the number of days that the result happened before (blue) or after (red) the query pattern.

2009, about the same day in 2010, and later in 2005, 2006, 2011 and 2012. For 2008, it happened one day earlier in 2009, about the same day in 2004, and later in the remaining years.

5. Conclusions and future work

Plant phenology studies are based on the analysis of several years of data. Average yearly plots of vegetation indexes are the preferred approach to evaluate phenological changes. Despite good results, the analysis based on average values is limited and can constraint the knowledge discovery process.

In this paper, we present PhenoVis, a framework for the visual phenological analysis of forest ecosystems. It contains the chronological percentage maps (CPM), a novel representation that is capable of discovering additional patterns by encoding percentage distributions of the data. We demonstrated CPM in a number of analysis scenarios, showing the additional insights that CPMs can bring to the analysis and how it can be used to identify species. The evaluation showed how automatic pattern searches can facilitate the detection of phenological singularities related to weather variations.

As future work, we intend to automatically detect phenological patterns. Currently this process is manual: the user informs the start and end dates of the query pattern. Automatic suggestions can improve the analysis using, for example, a box-plot [45] of the distribution. Another possibility is to use more images per day, decreasing lighting variations and artifacts. We also plan to investigate the integration of the CPM representation with machine learning techniques [4,7] to perform automatic species identification.

Acknowledgments

We thank the Phenological Eyes Network (PEN) for giving authorisation to use the TKY dataset (http://www.pheno-eye.org). We thank Shin Nagai and Kenlo Nishida Nasahara for the revision of this manuscript and suggestions to improve the final format. This research was supported by the São Paulo Research Foundation FAPESP and Microsoft Research Virtual Institute (grants #2010/52113-5, #2013/50169-1, and #2013/50155-0). 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. Also, we have benefited from funds given by CAPES, CNPq (grants 476685/2012-5, 309483/2011-5, 308851/2015-3), and FAPESP (grants #2007/52015-0, #2007/59779-6, #2009/18438-7, #2010/51307-0, and #2009/54208-6).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ins.2016.08.052.

References

[1] W. Aigner, S. Miksch, H. Schumann, C. Tominski, Visualization of time-oriented data, Springer Science & Business Media, 2011.
[2] D. Albers, C. Dewey, M. Gleicher, Sequence surveyor: Leveraging overview for scalable genomic alignment visualization, Visual. Comput. Graphics, IEEE Trans. 17 (12) (2011) 2392–2401.
[3] B. Alberton, J. Almeida, R. Henneken, R.S. Torres, A. Menzel, L.P.C. Morellato, Using phenological cameras to track the green up in a cerrado savanna and its on-the-ground validation, Ecol. Inf. 19 (2014) 62–70.
[4] J. Almeida, J.A. dos Santos, B. Alberton, R. da S. Torres, L.P.C. Morellato, Applying machine learning based on multiscale classifiers to detect remote phenology patterns in cerrado savanna trees, Ecol. Inf. 23 (0) (2014) 49–61. Special Issue on Multimedia in Ecology and Environment
[5] J. Almeida, J.A. Santos, B. Alberton, L.P.C. Morellato, R.S. Torres, Plant species identification with phenological visual rhythms, in: IEEE International Conference on eScience (eScience'13), 2013, pp. 148–154.
[6] J. Almeida, J.A. Santos, B. Alberton, L.P.C. Morellato, R.S. Torres, Visual rhythm-based time series analysis for phenology studies, in: IEEE International Conference on Image Processing (ICIP'13), 2013, pp. 4412–4416.
[7] J. Almeida, J.A. Santos, W.O. Miranda, B. Alberton, L.P.C. Morellato, R.S. Torres, Deriving vegetation indices for phenology analysis using genetic programming, Ecol. Inf. 26 (2015) 61–69.
[8] G. Andrienko, N. Andrienko, Spatio-temporal aggregation for visual analysis of movements, in: Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on, 2008, pp. 51–58.
[9] I. Boyandin, E. Bertini, P. Bak, D. Lalanne, Flowstrates: An approach for visual exploration of temporal origin-destination data, Comput. Graphics Forum 30 (3) (2011) 971–980.
[10] E. Bradley, D. Roberts, C. Still, Design of an image analysis website for phenological and meteorological monitoring, Environ. Modell. Software 25 (1) (2010) 107–116.
[11] L. Byron, M. Wattenberg, Stacked graphs – geometry & aesthetics, IEEE Trans. Visual. Comput. Graphics 14 (6) (2008) 1245–1252, doi:10.1109/TVCG.2008.166.
[12] W. Cui, X. Wang, S. Liu, N. Riche, T. Madhyastha, K.L. Ma, B. Guo, Let it flow: A static method for exploring dynamic graphs, in: Pacific Visualization Symposium (PacificVis), 2014 IEEE, 2014, pp. 121–128.
[13] H. Eerens, D. Haesen, F. Rembold, F. Urbano, C. Tote, L. Bydekerke, Image time series processing for agriculture monitoring, Environ. Modell. Software 53 (2014) 154–162.
[14] K. Fung, Numbers Rule Your World: The Hidden Influence of Probabilities and Statistics on Everything You Do, McGraw-Hill Education, 2010.
[15] A.R. Gillespie, A.B. Kahle, R.E. Walker, Color enhancement of highly correlated images. I. Decorrelation and HSI contrast stretches, Remote Sens. Environ. 20 (3) (1986) 209–235.
[16] J.A. Granados, E.A. Graham, P. Bonnet, E.M. Yuen, M.P. Hamilton, Ecoip: An open source image analysis toolkit to identify different stages of plant phenology for multiple species with pan-tilt-zoom cameras, Ecol. Inf. 15 (2013) 58–65.
[17] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, M. Streit, Lineup: Visual analysis of multi-attribute rankings, Visual. Comput. Graphics, IEEE Trans. 19 (12) (2013) 2277–2286.
[18] S. Hadlak, H. Schumann, C. Cap, T. Wollenberg, Supporting the visual analysis of dynamic networks by clustering associated temporal attributes, Visual. Comput. Graphics, IEEE Trans. 19 (12) (2013) 2267–2276.
[19] S. Havre, B. Hetzler, L. Nowell, Themeriver: visualizing theme changes over time, in: Information Visualization, 2000. InfoVis 2000. IEEE Symposium on, 2000, pp. 115–123, doi:10.1109/INFVIS.2000.885098.
[20] S. Havre, E. Hetzler, P. Whitney, L. Nowell, Themeriver: visualizing thematic changes in large document collections, IEEE Trans. Visual. Comput. Graphics 8 (1) (2002) 9–20, doi:10.1109/2945.981848.
[21] R. Ide, H. Oguma, Use of digital cameras for phenological observations., Ecol. Inf. 5 (5) (2010) 339–347.
[22] T. Inoue, S. Nagai, T.M. Saitoh, H. Muraoka, K.N. Nasahara, H. Koizumi, Detection of the different characteristics of year-to-year variation in foliage phenology among deciduous broad-leaved tree species by using daily continuous canopy surface images, Ecol. Inf. 22 (2014) 58–68. http://dx.doi.org/10.1016/j.ecoinf.2014.05.009.
[23] A. Ito, N. Saigusa, S. Murayama, S. Yamamoto, Modeling of gross and net carbon dioxide exchange over a cool-temperate deciduous broad-leaved forest in japan: Analysis of seasonal and interannual change, Agric. For. Meteorol. 134 (1-4) (2005) 122–134. http://dx.doi.org/10.1016/j.agrformet.2005.11.002. Long-Term Carbon Exchange at the Takayama, Japan Forest
[24] D.A. Keim, Designing pixel-oriented visualization techniques: Theory and a.0pplications, IEEE Trans. Visual. Comput. Graphics 6 (1) (2000) 59–78.
[25] D.A. Keim, Information visualization and visual data mining, IEEE Trans. on Vis. and Comp. Graph. 8 (1) (2002).
[26] D.A. Keim, M.C. Hao, U. Dayal, M. Hsu, Pixel bar charts: A visualization technique for very large multi-attribute data sets, Inf. Visual. 1 (1) (2002) 20–34.
[27] J. Li, K. Zhang, Z.-P. Meng, Vismate: Interactive visual analysis of station-based observation data on climate changes, in: Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on, 2014, pp. 133–142.
[28] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, J. Heer, Selecting semantically-resonant colors for data visualization, Comput. Graphics Forum 32 (3pt4) (2013) 401–410.
[29] J.T. Morisette, A.D. Richardson, A.K. Knapp, J.I. Fisher, E.A. Graham, J. Abatzoglou, B.E. Wilson, D.D. Breshears, G.M. Henebry, J.M. Hanes, L. Liang, Tracking the rhythm of the seasons in the face of global change: phenological research in the 21st century, Front. Ecol. Environ. 7 (5) (2009) 253–260.
[30] S. Nagai, T. Maeda, M. Gamo, H. Muraoka, R. Suzuki, K.N. Nasahara, Using digital camera images to detect canopy condition of deciduous broad-leaved trees, Plant Ecology & Diversity 4 (1) (2011) 79–89.
[31] K.N. Nasahara, S. Nagai, Review: Development of an in situ observation network for terrestrial ecological remote sensing: the phenological eyes network (PEN), Ecol. Res. 30 (2) (2015) 211–223.
[32] G. Negi, Leaf and bud demography and shoot growth in evergreen and deciduous trees of central himalaya, india, Trees 20 (4) (2006) 416–429.
[33] C.W. Ngo, T.C. Pong, R.T. Chin, Detection of gradual transitions through temporal slice analysis, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'99), 1999, pp. 1036–1041.
[34] K.T. Nguyen, T. Ropinski, Large-scale multiple sequence alignment visualization through gradient vector flow analysis, in: Biological Data Visualization (BioVis), 2013 IEEE Symposium on, 2013, pp. 9–16.
[35] K. Nishida, Phenological eyes network (PEN) - a validation network for remote sensing of the terrestrial ecosystems., AsiaFlux Newslett. 21 (2007) 9 13.
[36] G. Oliveira, J. Comba, R. Torchelsen, M. Padilha, C. Silva, Visualizing running races through the multivariate time-series of multiple runners, 2012 25th SIBGRAPI Conf. Graphics, Patterns Images 0 (2013) 99–106.
[37] C. Reas, B. Fry, Processing: A Programming Handbook for Visual Designers and Artists, MIT Press, 2014.
[38] A.D. Richardson, B.H. Braswell, D.Y. Hollinger, J.P. Jenkins, S.V. Ollinger, Near-surface remote sensing of spatial and temporal variation in canopy phenology, Ecol. Appl. 19 (6) (2009) 1417–1428.
[39] M.D. Schwartz, Phenology: an integrative environmental science, 2nd, Springer, 2013.

[40] O. Sonnentag, K. Hufkens, C. Teshera-Sterne, A.M. Young, M. Friedl, B.H. Braswell, T. Milliman, J. O Keefe, A.D. Richardson, Digital repeat photography for phenological research in forest ecosystems, Agric. For. Meteorol. 152 (2012).
[41] T. Udelhoven, Timestats: A software tool for the retrieval of temporal patterns from global satellite archives, IEEE J. Selected Topics Appl. Earth Observ. Remote Sens. 4 (2) (2011) 310–317.
[42] G.-R. Walther, E. Post, P. Convey, A. Menzel, C. Parmesan, T.J.C. Beebee, J.-M. Fromentin, O. Hoegh-Guldberg, F. Bairlein, Ecological responses to recent climate change, Nature 416 (6879) (2002) 389–395.
[43] Z. Wang, M. Lu, X. Yuan, J. Zhang, H. Van De Wetering, Visual traffic jam analysis based on trajectory data, Visual. Comput. Graphics, IEEE Trans. 19 (12) (2013) 2159–2168.
[44] C. Ware, Information Visualization: Perception for Design, 3, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2012.
[45] D.F. Williamson, R.A. Parker, J.S. Kendrick, The box plot: a simple visual method to interpret data, Ann. internal med. 110 (11) (1989) 916–921.
[46] J. Wood, D. Badawood, J. Dykes, A. Slingsby, Ballotmaps: Detecting name bias in alphabetically ordered ballot papers., IEEE Trans. Vis. Comput. Graph. 17 (12) (2011) 2384–2391.
[47] X. Zhang, M.A. Friedl, B. Tan, M.D. Goldberg, Y. Yu, Long-term detection of global vegetation phenology from satellite instruments, Phenol. Climate Change (2012) 297–320.