UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ROGER ALMEIDA LEITE

PhenoVis: a visual analysis tool to phenological phenomena

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Prof. Dr. João Luiz Dihl Comba
Advisor

Prof. Dr. Lucas Mello Schnorr
Coadvisor

Porto Alegre,RS
2015

CIP – CATALOGING-IN-PUBLICATION

Almeida Leite, Roger

PhenoVis: a visual analysis tool to phenological phenomena / Roger Almeida Leite. – Porto Alegre,RS: PPGC da UFRGS,

.

66 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre,RS, BR–

,

. Advisor: João Luiz Dihl Comba; Coadvisor: Lucas Mello Schnorr.

1. Visual Analytics. 2. Multidimensional Analysis. 3. Percentage Distribution. 4. Similarity Ranking. 5. Phenology. I. Dihl Comba, João Luiz. II. Mello Schnorr, Lucas. III. Title.

“The purpose of information visualization is to amplify cognitive performance, not just to create interesting pictures. Information visualizations should do for the mind what automobiles do for the feet.”

— CARD, 2008, P. 539

ABSTRACT

Phenology studies recurrent periodic phenomena of plants and their relationship to environmental conditions. Monitoring forest ecosystems using digital cameras allows the study of several phenological events, such as leaf expansion or leaf fall. Since phenological phenomena are cyclic, the comparative analysis of successive years is capable of identifying interesting variation on annual patterns. However, the number of images collected rapidly gets significant since the goal is to compare data from several years. Instead of performing the analysis over images, experts prefer to use derived statistics (such as average values). We propose PhenoVis, a visual analytics tool that provides insightful ways to analyze phenological data. The main idea behind PhenoVis is the Chronological Percentage Maps (CPMs), a visual mapping that offers a summary view of one year of phenological data. CPMs are highly customizable, encoding more information about the images using a pre-defined histogram, a mapping function that translates histogram values into colors, and a normalized stacked bar chart to display the results. PhenoVis supports different color encodings, visual pattern analysis over CPMs, and similarity searches that rank vegetation patterns found at various time periods. Results for datasets comprising data of up to nine consecutive years show that PhenoVis is capable of finding relevant phenological patterns along time.

Keywords: Visual Analytics, Multidimensional Analysis, Percentage Distribution, Similarity Ranking, Phenology.

PhenoVis: uma ferramenta de análise visual para fenômenos fenológicos

RESUMO

Fenologia estuda os fenômenos recorrentes e periódicos que ocorrem com as plantas. Estes podem vir a ser relacionados com as condições ambientais. O monitoramento de florestas, através de câmeras, permite o estudo de eventos fenológicos como o crescimento e queda de folhas. Uma vez que os fenômenos fenológicos são cíclicos, análises comparativas de anos sucessivos podem identificar variações interessantes no comportamento destes. No entanto, o número de imagens cresce rapidamente para que sejam comparadas lado a lado. PhenoVis é uma ferramenta para análise visual que apresenta formas para analisar dados fenológicos através de comparações estatísticas (preferência dos especialistas) derivadas dos valores dos pixels destas imagens. A principal ideia por trás de PhenoVis são os mapas percentuais cronológicos (CPMs), um mapeamento visual com uma visão resumida de um período de um ano de dados fenológicos. CPMs são personalizáveis e conseguem representar mais informações sobre as imagens do que um gráfico de linha comum. Isto é possível pois o processo envolve o uso de histogramas pré-definidos, um mapeamento que transforma valores em cores e um empilhamento dos mapas de percentagem que visa a criação da CPM. PhenoVis suporta diferentes codificações de cores e análises de padrão visual sobre as CPMs. Pesquisas de similaridade ranqueiam padrões parecidos encontrados nos diferentes anos. Dados de até nove anos consecutivos mostram que PhenoVis é capaz de encontrar padrões fenológicos relevantes ao longo do tempo.

Palavras-chave: Análise Visual, Análise multidimensional, Distribuição de Percentagem, Ranking de Similaridade, Fenologia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Phenology is the study of any periodic event presented in nature. It can involve animals as well as plants. Our focus in this work is the better understanding of phenological data. In order to do that, our main efforts have been done in mining and representing recurrent phenological phenomena of plants and their relationship to environmental conditions (PHENOLOGICAL DATA, NETWORKS, AND RESEARCH: NORTH AMERICA, 2013). This study is critical to analyze the impact of environmental changes on vegetation and ecosystem processes and long-term impacts of climate change (NEGI, 2006; WALTHER et al., 2002; MORISETTE et al., 2009; PHENOLOGICAL DATA, NETWORKS, AND RESEARCH: NORTH AMERICA, 2013). Examples of cyclic phenomena in plants are flowering and fruiting phase.

The remote monitoring of forest or vegetation areas using cameras followed by subsequent analysis has been shown to be a very promising approach to study plant behavior (RICHARDSON et al., 2009; NAGAI et al., 2011; SONNENTAG et al., 2012). In this scenario, cameras capture daily pictures from a specific viewpoint and time of the day. The variations of these images over time may identify interesting phenological events. One way, to capture variations, is to use simple statistics such as average plots of derived quantities (e.g. green component of each pixel). However, as it is well known (FUNG, 2010), using the average for analysis can often elude interesting patterns hidden in the data.

Currently, the full amount of information that are generated by the monitoring system is not utilized to its full potential. For instance, solutions (RICHARDSON et al., 2009; SONNENTAG et al., 2012) to this problem are insufficient because two different images might have the same average score (current context detailed explained in Section 1.1). Another problem is the lack of automation in the search for regular patterns common in plant behavior. Current research in the field uses analyst intuition to find patterns of plant behavior. Nevertheless, this interactive approach is not scalable, due to the size of data to analyze, and error-prone, because of human decisions.

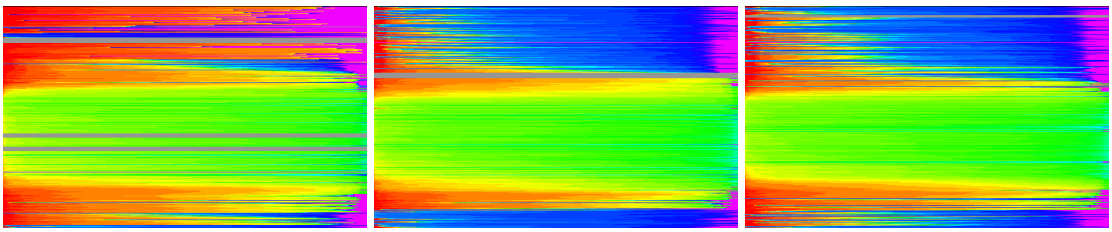
In this work we introduce PhenoVis, a visual analytics tool that aims at providing insightful ways to analyze phenological data. The main idea behind PhenoVis is the notion of a Chronological Percentage Map (CPM), a visual mapping that combines derived distributions from all images of a given year to create normalized stacked bar charts. PhenoVis uses the additional information encoded to support similarity searches, useful for comparing data from different years. It also provides a customizable multi-rank comparison of data from different years, with filters for specific periods within a year or sub-regions associated with given plant species. In summary, the main contributions introduced are:

- PhenoVis, a visual analytics tool to perform a comparative analysis of phenological data of multiple years;
- The design of CPMs, which combines a more expressive representation of phenological data using percentage distributions and a visual expression of this information using color-coded normalized stacked bar charts;
- Similarity algorithms that allow to search for similar phenological patterns across years, either happening at a fixed or moving window of time;

- A customizable ranking comparison of years, with filters that allow selections of specific time periods or regions;
- Intraspecies and interspecies differences behavior identification;
- Non-identified areas species estimation;
- Case studies that validate PhenoVis in phenological analysis tasks, such as pattern and outlier identification.

Some results are presented by Figure 1.1 in order to preview part of the PhenoVis features. This figure shows the normalized stacked bar charts for years 2004, 2005 and 2006. Images are captured daily at noon, leading to 365 images per year. A percentage distribution is built for each image from a derived quantity computed at each pixel (e.g. hue variation). Distributions are mapped into a normalized stacked bar chart using a predefined color scale, here displayed vertically, with each line corresponding to one day. Such visual mapping is called a Chronological Percentage Map (CPM). Missing data is show as gray rows, clearly visible in 2004. Green colors identify the leafing period (visible in all years) and red colors with warmer or brighter pixels. Note how 2004 differs from 2005 and 2006, which was due to higher temperatures and humidity indexes. Following this CPMs, we are able to observe that 2004 is very different from the others. One explanation is given by meteorological data collected for these years. The year of 2004 had higher temperatures and was more humid than the others.

Figure 1.1: PhenoVis analysis for 2004, 2005 and 2006.



(a) 2004 (b) 2005 (c) 2006

Source: Author.

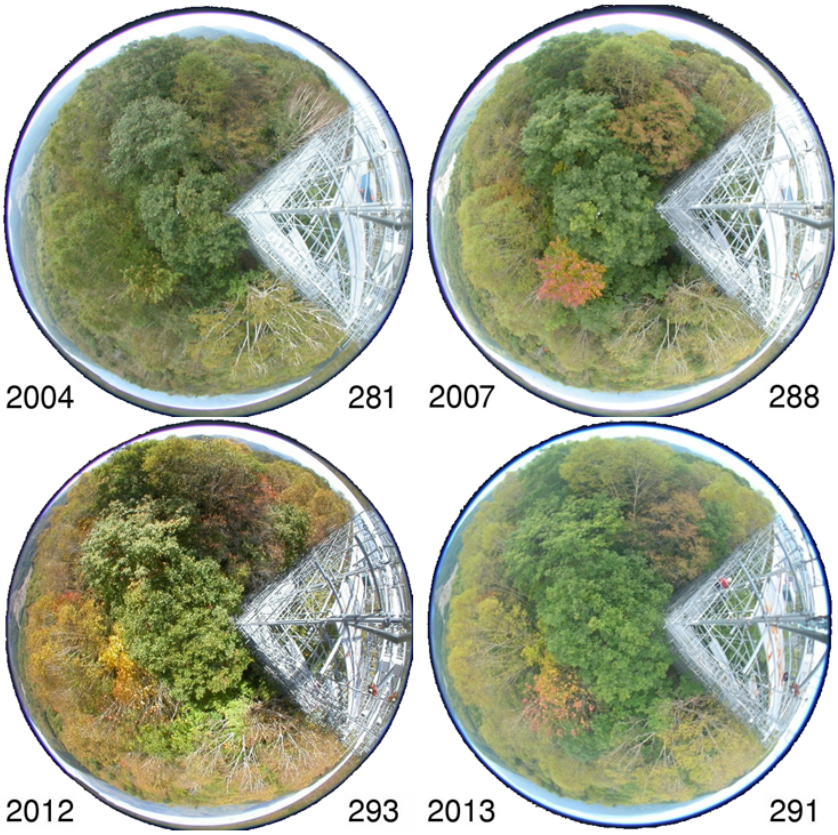
1.1 Limitations on Using the Average

Biologists analyze plant phenology data over time through vegetation indexes averages. In this section, we analyze the shortcomings of using average scores in phenology studies. Since the datasets are big and constantly increasing, the large amount of pictures (at least one for each day) poses many difficulties in this analysis. The comparative analysis of successive years is largely used to identify interesting patterns revealed by the variation of the vegetation index. However, one of the challenges refers to choose a visualization technique for this data; keeping as much information as possible.

In the literature, each image is transformed into a single vegetation index value. For instance, g_{cc} (see Section 2.2), that represents the average of the green intensity considering all the relevant pixels of that image. This average of green intensity is a measure widely used within the phenology community to better understand phenological phenomena. To display this information, biologists draw for each year a 2D-plot of the average

of green intensity in relation to the days of the year (time). With this common approach, they are capable of studying the phenological development for many years and are able to detect, for example, patterns of plant grow or even the impact of climate change on specific types of vegetation.

Figure 1.2: Four images of the same region from different years (2004, 2007, 2012, and 2013) and days (281, 288, 293, and 291). Although the images look very different, they have the same average $g_{cc} = 0.3905$.



Source: (NAGAI et al., 2011).

While widely accepted, the usage of average as a comparison vegetation index between images can potentially hide important behavior and eventually leads to wrong conclusions. Figure 1.2 illustrates such situation considering four different images captured by the same device in different years and days. As we can see, each image looks very different when compared to the others, but all of them share the exactly same average of green intensity $g_{cc} = 0.3905$ (considering that the metallic structure is put aside). Therefore, using the traditional average-based approach, phenologists will be lead to conclude that all these four images are the same according to the metric, even if they are completely different in reality.

1.2 Dissertation Structure

The dissertation is organized as follows. In Chapter 2 so Chapter 3 discuss related work in the field of Phenology and how our approach differs from existing solutions. It also presents visualization designs and techniques that are similar to CPM. PhenoVis is described in Chapter 4 the chronological percentage map design is presented. Whereas

in Chapter 5, similarity search, and ranking algorithms are explained. More details about the prototype are given in Chapter 6. Chapter 7 lists the results obtained with PhenoVis, using five different datasets. A brief discussion and the future work is described at the end of the thesis. Discussions and limitations are shown in Chapter 8, whereas the conclusion and possible future works are given in Chapter 9.

2 BACKGROUND ON PLANT PHENOLOGY

In this Chapter, we summarize essential concepts in plant phenology, and list fundamental questions raised in the analysis of this data.

Leafing is the plant phenological event that defines the growth season and controls crucial ecosystems processes such as nutrient cycling, water storage, and carbon balance, regulating the productivity in terrestrial ecosystems and the dynamics of carbon sequestration (REICH, 1995; ROTZER; GROTE; PRETZSCH, 2004; NEGI, 2006). Temperature is regarded as the main cue for leaf out in temperate plants. One of the central questions on leaf phenology of temperate vegetation is: in which rate, changes on spring temperature may lead to shifts in the start of the leafing season? This question is of crucial importance since there is a general increase in temperature due to global warming, and those changes are related to an advance on leaf phenology across temperate regions (MENZEL et al., 2006; WALTHER et al., 2002). Therefore, actual research is searching for more refined ways to detect the starting of leaf season and shifts related to changes in spring temperature.

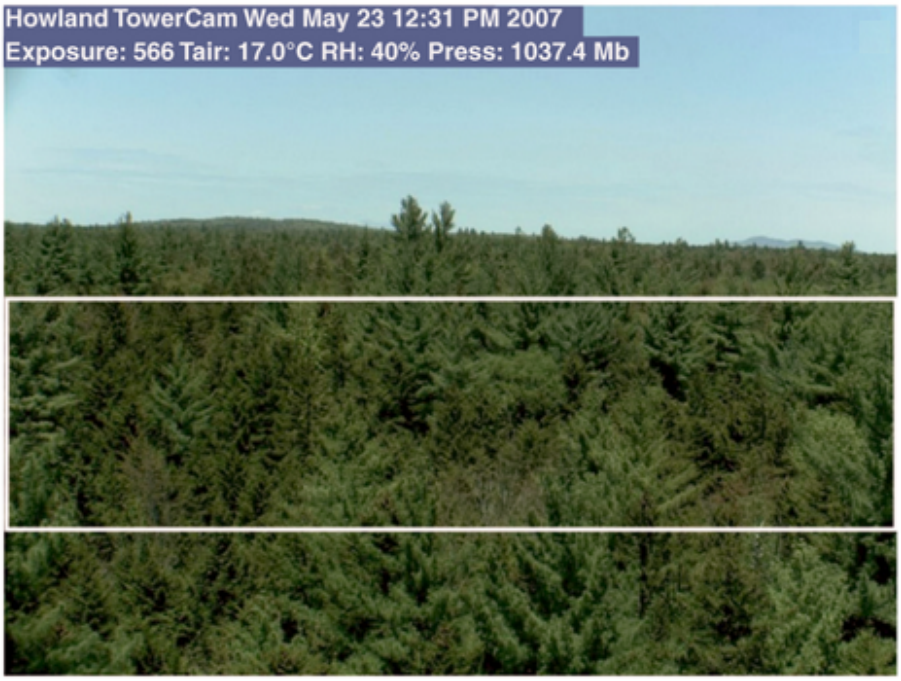
Many works have focused in near-surface monitoring systems (CRIMMINS; CRIMMINS, 2008; RICHARDSON et al., 2007; POLGAR; PRIMACK, 2011). Near remote phenology with digital cameras offers daily information on leaf changes – a more refined plant analysis data is obtained in researches (MENZEL et al., 2006; POLGAR; PRIMACK, 2011) . Leaf fall or leaf loss is also the focus of several studies. We need new tools to describe those patterns and define the start of the leafing season as its changes over time. Finally, one basic challenge for phenology researchers is the plant species identification within digital images (ALMEIDA et al., 2014). The CPM representation is a promising alternative to the task of identifying similar species across seasons, a step forward for phenology research.

2.1 Plant Phenology Studies Using Digital Cameras

Data analysis from remote monitoring of plants by digital cameras have demonstrated its applicability for obtaining information of phenological events and as a solution for continuous, unbiased monitoring of leaf changes (RICHARDSON et al., 2007, 2009; AHRENDS et al., 2009). The current state-of-the-art in this field computes different vegetation indexes (see Section 2.2) from a region of the captured image, and displays it in a line graph (RICHARDSON et al., 2009; SONNENTAG et al., 2012; ALBERTON et al., 2014). Line graphs aggregates information, and this characteristic can lead to simplified interpretation. This approach ignores some features of the data that might be significant for the analysis of the results (SCHNORR; LEGRAND, 2013). In (ALMEIDA et al., 2012), we can find phenology studies based on vegetation images using different computational features, such as machine learn (specifically weak learners technique), in order to better understand Cerrado Forest phenomena, Brazil.

For instance, a temperate forest has been monitored in Takayama, Japan, with the aim of identifying leaf color change patterns among three local dominant plant species throughout the years of 2005 to 2007 (NAGAI et al., 2011). That study concluded that differences among species can be detected using normalized RGB values. However, the

Figure 2.1: Howland forest. This is an example of image(data) from a remote monitored forest. White lines define the region of interest concept, presented by (RICHARDSON et al., 2009).



Source: (RICHARDSON et al., 2009).

method was not able to identify differences between plant features during the spring period, suggesting the need for a more detailed approach to compare leaf pigments color, as chlorophylls and anthocyanins, with RGB values. Therefore, new features or visualization techniques are necessary to allow an accurate differentiation of species within seasons or specific time periods (e.g., leaf season) over the years. The proposed CPM representation is a step forward in this direction.

2.2 Basic Vegetation Index Concept

Leafing phenology is observed using digital images for extracting and quantifying the green channel out of the RGB (Red, Green and Blue) channels. The normalized *RGB* chromatic coordinate (*RGBcc*) index is referred as the most suitable index to detect leaf color changes and the most efficient to suppress light variation (GILLESPIE; KAHLE; WALKER, 1987; M. et al., 1995; SONNENTAG et al., 2012). The *RGBcc* undergoes a nonlinear transform, as follows:

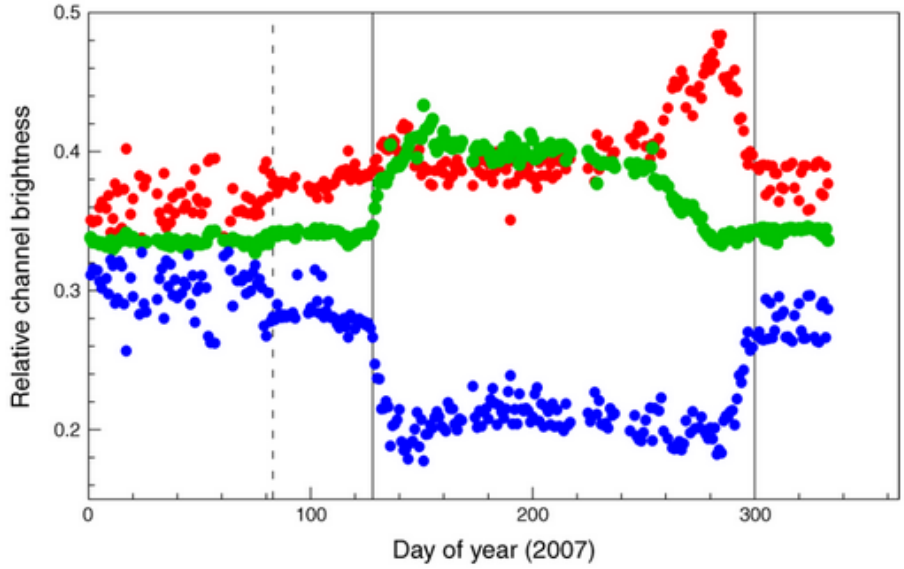
$$r_{cc} = \frac{R}{(R+G+B)}; \quad g_{cc} = \frac{G}{(R+G+B)}; \quad b_{cc} = \frac{B}{(R+G+B)}; \tag{2.1}$$

$$r_{ex} = (R*2) - (G+B); \quad g_{ex} = (G*2) - (R+B); \quad b_{ex} = (B*2) - (R+G); \tag{2.2}$$

We use the pixels *RGB* values to obtain information that are semantically relevant in biology phenomena (RICHARDSON et al., 2009). These colors channels reflect the leaf pigmentation. For example, when the green average increases, represents the Chlorophyll levels climb during leaf expansion phase. Another example is Leaf fall cycle that can

be analysed by the *Xanthophylls* (an yellow pigment) and *Anthocyanins* (normally red pigment) levels (LEE et al., 2003). Among these equations 2.1, the g_{cc} is the most used in phenological analysis and is one of the main vegetation index used to present results in PhenoVis. See in Figure 2.2 a vegetation index plot presented by (RICHARDSON et al., 2009).

Figure 2.2: Vegetation index plot using r_{cc} (red points), g_{cc} (green points), and b_{cc} (blue points).



Source: (RICHARDSON et al., 2009).

Vegetation index like these can be used to each pixel of an image sample. Usually it is calculated to all target pixels and, after that, an average off all results is defined to represent the image sample. However, this technique aggregates information.

2.3 Phenological Analysis

The increasing accessibility to phenological data with high spatio temporal resolution has enabled a detailed analysis of vegetation properties. At the same time, it raises the need for appropriate visualization techniques to capture such properties and to support the mining of this type of dataset. Most existing works in phenology have focused on primary visualization strategies such as display a 2D graph plotting the computed vegetation index against time.

Bradley (BRADLEY; ROBERTS; STILL, 2010) created a web-based interface to provide different display options for archived image review and to chart time-series obtained from phenological and meteorological observations. TimeStats (UDELHOVEN, 2011) is a free software that offers tools for visualization of long-term remote sensing data archives, such as parametric and non-parametric methods for trend detection, linear regression, and frequency analysis. EcoIP (GRANADOS et al., 2013), a toolkit to estimate an onset and ending dates of phenophases of plant species relies on a Naive Bayesian model created from a set of training images, which is used to provide temporal estimators. A variety of color transformations is used to adjust the accuracy of the estimations. Another recent trend refers to the construction of toolboxes. Eerens (EERENS et al., 2014),

for instance, developed SPIRITS, a stand-alone toolbox to produce evidence-based information for crop production analysts. It includes a large number of features to analyze image time series and to create maps and graphs for vegetation status analysis.

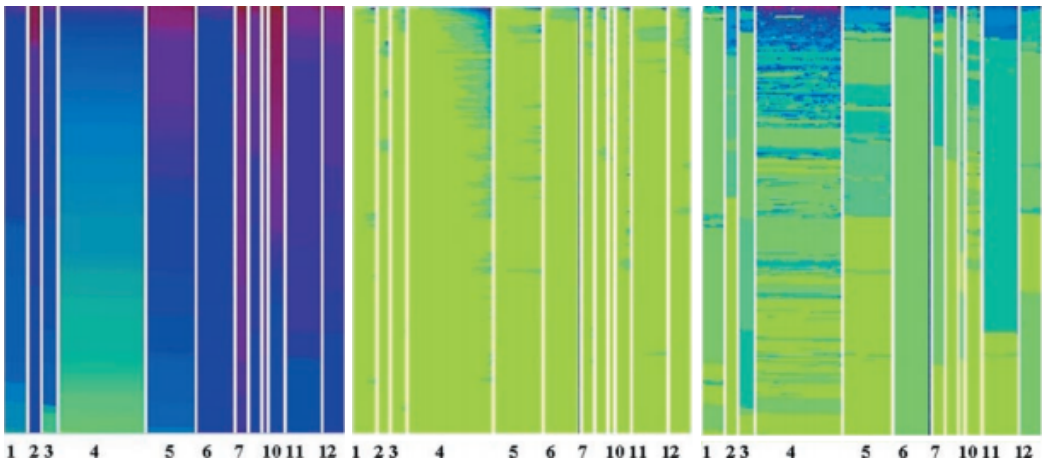
Different from those initiatives, our proposal relies on visual encodings designed to capture distribution aspects from vegetation data. A similar approach is taken in (ALMEIDA et al., 2013a,b), which presented different strategies for encoding image time series as visual rhythms (NGO; PONG; CHIN, 1999). Such representations have proven to be a powerful tool for distinguishing the behavior of various 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 RELATED WORK

In this Chapter we review related visual analytics and information visualization techniques, as well as approaches to analyse phenological data extracted from images.

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. Keim (KEIM, 2002) propose a classification of information visualization and visual data mining techniques based on data types. The main visual mining benefit from this work is the use of the combination from human perception capabilities and the computer storage capacity and power. Rectangular regions such as matrices are a common alternative to present aggregated data to allow identification of patterns. Pixel bar charts (KEIM et al., 2002) extends traditional bar charts by coloring pixels inside bars with derived information from data attributes (see Figure 3.1). Since his technique display more information than a regular bar chart, it allows large datasets to be visualized with more efficiency. The visual presentation of data follows a pixel-placement ordering, which was demonstrated to reveals patterns in data, as well as serve the basis for data mining queries. The underlying motivation behind pixel bars was an inspiration when designing PhenoVis, in particular the CPM notion and the ability to search for similar patterns in CPMs from different years.

Figure 3.1: general-pixel bar charts. Analysing each bar, y-axis is the number of consumers (normalized in order to fit equally and allow comparison among the samples), while product type is represented by x-axis. In left to right, the colors in graphs are: Amount spent(a), number of visits(b), and sales quantity(c).

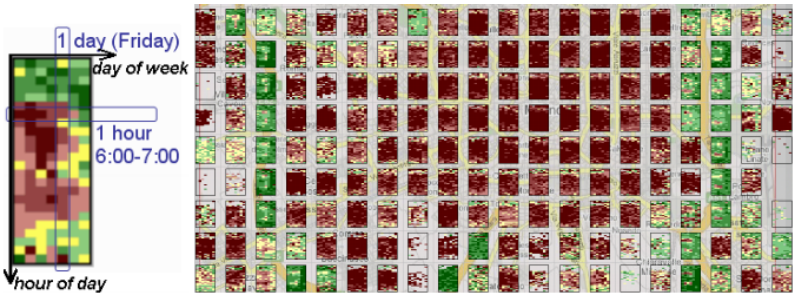


(a) (b) (c)

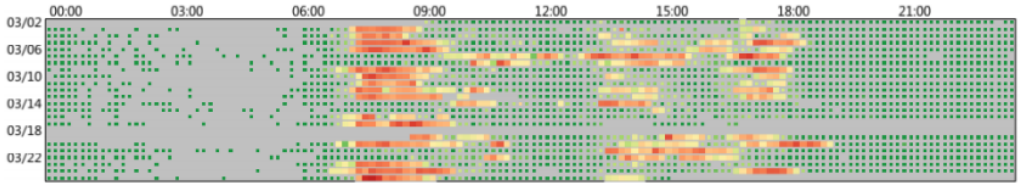
Source: (KEIM et al., 2002).

There are several works that use matrix-encoded information to perform visual analysis. Andrienko (ANDRIENKO; ANDRIENKO, 2008) uses stacked bar charts to display temporal variation of speed traffic data. The normalized stacked bar charts used in CPMs resemble the images showed in their work, especially when using categorical color mappings. In Figure 3.2 (a), it is interesting to observe the different curves that each part of the classified and sort data generates when arranged in relation to time. The analysis of visual traffic also appears displayed as pixel oriented map in (WANG et al., 2013), with trajectory information chronologically encoded in a matrix. Relevant events are highlighted by this technique and we can verify the importance of a the palette choice (see Figure 3.2 (b)). A tool that allows interactive data exploration have also been developed in this work. In (OLIVEIRA et al., 2013), a pixel matrix, Figure 3.2 (c), encoding heart-rate information is used to identify unusual patterns during a running race. Three visualization designs are presented. These were motivated by answer questions formulated by an expert.

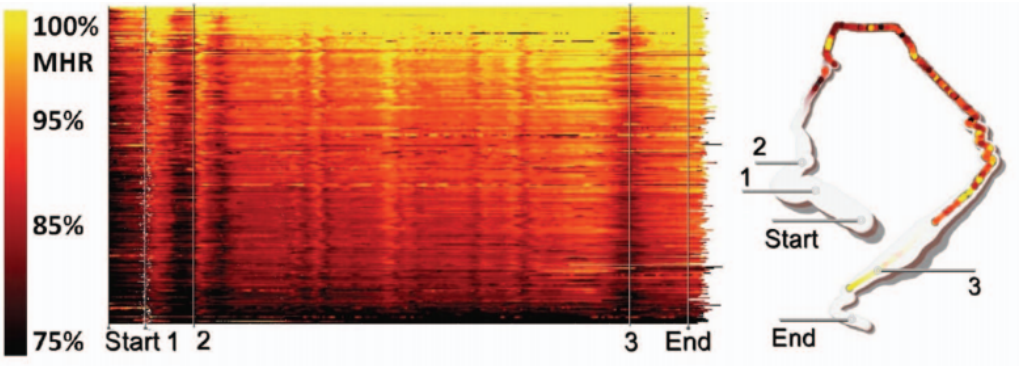
Figure 3.2: Image (a) shows the variation of the median speeds in relation to the days of the week. In the diagram, columns represent days, rows represent hours. Speed values color map the cells from the diagram. Shades of red represent low speeds, on the other hand shades of green are high speeds. Image (b) displays a traffic congestion pattern by analysing the red shades. This pattern appears in relation to the hour of the day. Image (c) allows the analysis of a running race trace with multiple runners in relation to their maximum hearth rate.



(a) [Source: (ANDRIENKO; ANDRIENKO, 2008).



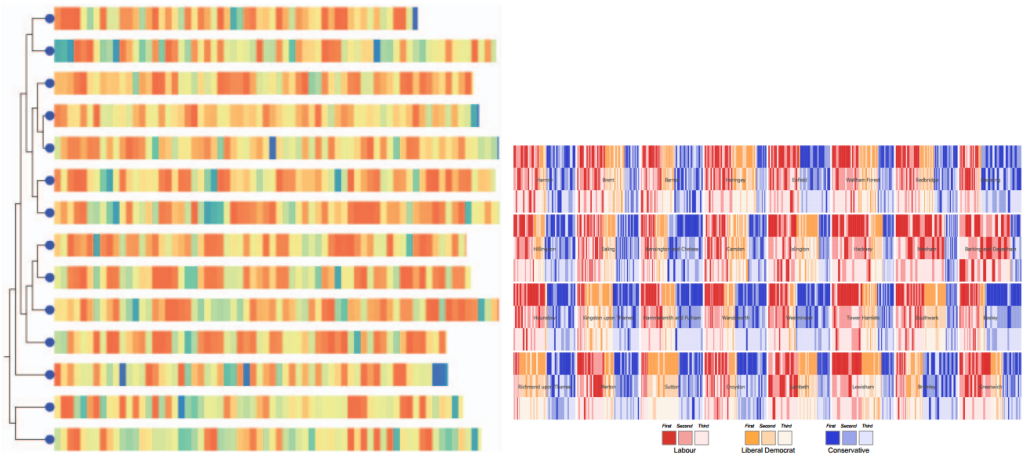
(b) Source: (WANG et al., 2013).



(c) Source: (OLIVEIRA et al., 2013).

Developed to compare genomic sequences, (ALBERS; DEWEY; GLEICHER, 2011) uses colored mapped matrices (see Figure 3.3 (a)) and can also be used to other applications such as text analysis. However the scalability is still limited since "vertical" aggregation strategies to group data is not presented. The of use matrices to compare genomic sequences is also described in (NGUYEN; ROPINSKI, 2013). Detected features are emphasized by the extracted information. Nonetheless visualizations with can be generated with different alignment algorithms and these can not be compared yet. BallotMaps (WOOD et al., 2011) encode politicians votes in matrices to identify voting patterns. As we see in Figure 3.3 (b), sort and color map data allow determine name influence degree and reveal preferences from each party. Flowstrates (BOYANDIN et al., 2011) encodes into a heatmap matrix several origin-destination data, re-ordering rows to reveal interesting patterns. No math technique is computed in order to reveal patterns in the matrices. This tool focus is the interactive exploration.

Figure 3.3: Image (a) displays 14 bacterias. These are colored by group frequency and show their spatial organization and the conservation of genes. Orange shades blocks reveals that genes are well conserved overall. Image (b) represents a graph divided by wards(bars) inside grid squares, these groups represent borough. This shows the vote popularity from each of the three main parties' politician candidates. This visualization is from 2010 local government elections.



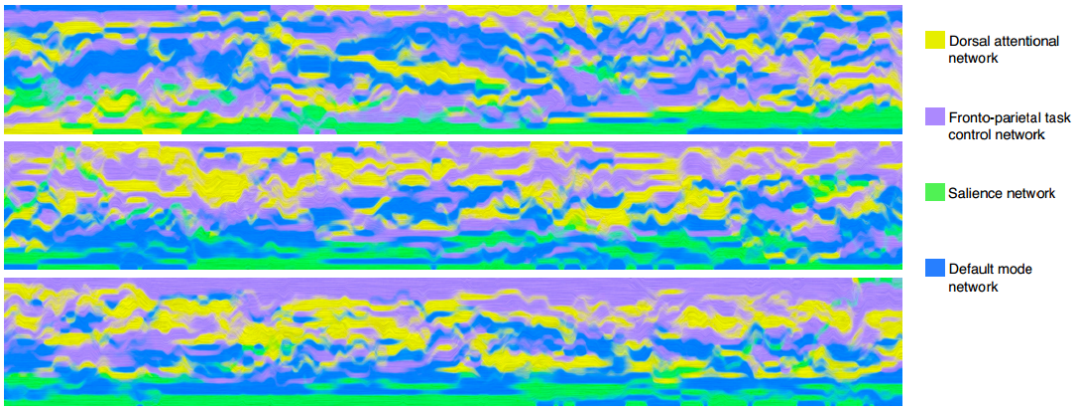
(a) Source: (ALBERS; DEWEY; GLEICHER, 2011).

(b) Source: (WOOD et al., 2011).

The exploration of dynamic graphs using matrix-encoded information is given in (CUI et al., 2014). Multiple graph metrics are allowed by the design presented by this article. On the other hand, the patterns appearance highly depends on the adopted metric. In (HADLAK et al., 2013), the visualization of dynamic networks are aggregated by clustering temporal attributes. Part of the proposed aggregation visualization is based on the events recurrence. This characteristic highlights meaningful behaviors without miss the events variation, see Figure 3.4.

Vismate (JIE LI KANG ZHANG, 2014) presents a climate change data comparison using a global radial map. Although it uses a radial representation, the compact representation resembles the CPM stacked bar charts. LineUp (GRATZL et al., 2013) describes different ways to present general-ranking attributes. In this work validation, users could solve complex ranking problems in a short period of time. This visualization and interactive technique (showed by Figure 3.5) allows a finer analysis of general attribute

Figure 3.4: In these three images, x axis represents time and the y axis represents the connectedness of four different functional regions. These are encoded by different colors. Top to bottom are respectively younger to older results



Source: (CUI et al., 2014).

rankings. The tool also enables interact balancing different heights to each attribute in order to compare elements from the same rank. The general-ranking visualization developed in our work shows the need for ranking visualization when performing comparative phenology analysis. For a comprehensive survey of other ways to perform visual analysis of time-oriented data, we refer the reader to (AIGNER et al., 2011).

Figure 3.5: Multi-attribute rankings.

Source: (GRATZL et al., 2013).

4 PHENOVIS – CHRONOLOGICAL PERCENTAGE MAP (CPM)

CPM is the main concept behind PhenoVis. After PhenoVis calculates a histogram for each image, where the value of each pixel is defined according to a user-defined metric, CPM represents the found behaviors. Since the technique involved in CPM keeps more information about complex data (such as images), the CPM representation gives more details about plant behavior than traditional average-based plots. However, it keeps the over-viewing capabilities necessary to detect patterns on data, either manually or automatically. This subsection details the steps to create a CPM. Its construction consists of six steps that transform a sequence of images into a normalized stacked bar chart, as illustrated in Figure 4.1.

4.1 CPM generation flow

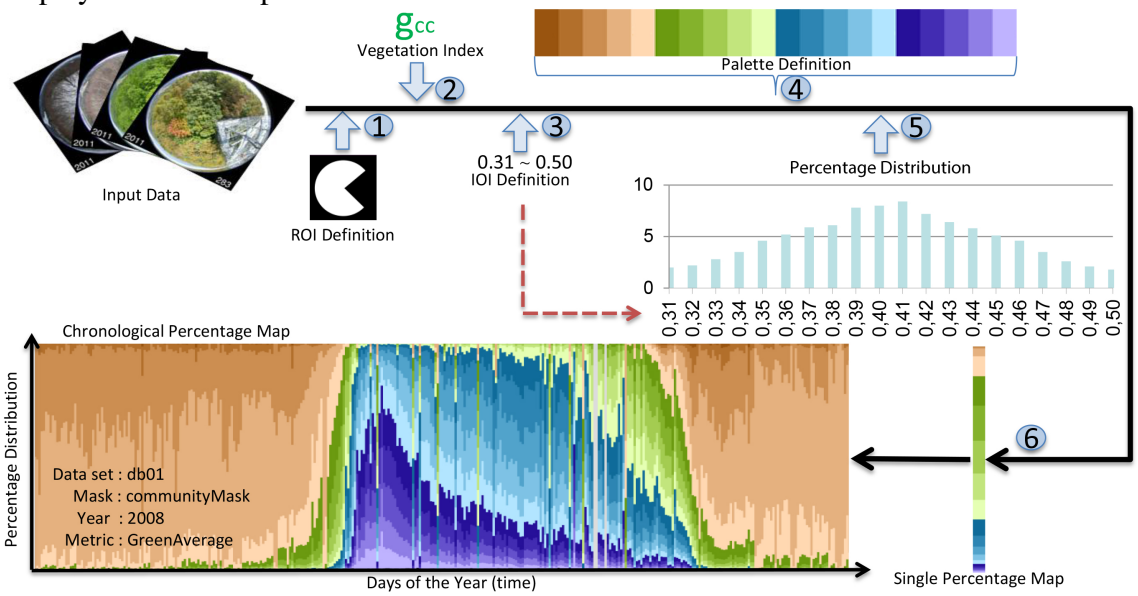
First, each input image is filtered to remove parts irrelevant to the analysis (step 1). The remaining pixels of the input image are used to compute derived attributes using a given metric (step 2). The range of values of the derived attribute can be specified to narrow the analysis into specific intervals of interest (step 3). The derived attributes will be used to compute a percentage histogram. The granularity of the histogram and the color mapping used for each histogram entry is defined in step 4. The percentage distribution for the selected interval of interest and selected metric is computed in step 5. The resulting distribution has information from a single image (one day). We create a normalized stacked bar chart (called a percentage map) for each percentage distribution using the color mapping defined. Since we take as input a set of images collected over time, we stack the percentage maps in chronological order (step 6). The resulting CPM can be created either in portrait or landscape orientations. After detail each of these steps in the order from our technique, the subsection ends debating about the data size influence followed by a brief discussion about the "multidimensional increasing" interpretation that is presented by the percentage map approach.

4.1.1 Step 1: Filtering by region of interest (ROI)

Phenologists have access to a series of images of forests or natural fields captured automatically by special cameras. Very often, they want to focus the phenological study in only a portion of this image. It can be a plant of a given species that is located in a part of the picture, but also a group of plants that share common characteristics. It can also be the whole picture if the biologist wants to study the community as a whole along time. These different examples illustrate the region of interest (ROI): a portion of the image that is under study. For each analysis, the first task of the analyst is to define the region of interest considering the set of images available.

Region of interest is defined by a mask image made or oriented by expert. It has the same size dimensions of the samples that are under study, but are composed exclusively of black and white pixels. The portions of the image where pixels are colored white are the region of interest defined by the analyst. In others words, a mask is a stencil image.

Figure 4.1: The Chronological Percentage Map construction comprises six steps: (Step 1) images are filtered by region of interest; (2) data transformation considering the phenology metric; (3) filtering by interval of interest; (4) choosing the data distribution granularity; (5) calculating the data distribution, and finally (6) the CPM stacked bar chart, displayed in landscape mode.



Source: Author.

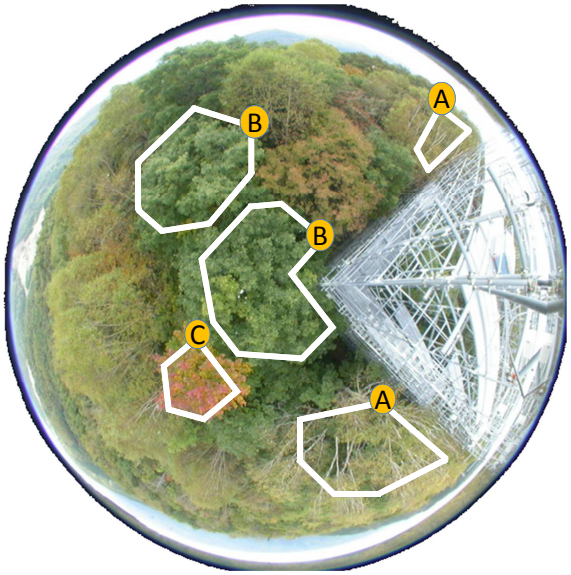
Besides being used to select parts of the picture that are of interest, the use of masks are beneficial because they allow the biologist to take out information that is considered completely irrelevant to the analysis. Examples include portions of the image that lacks vegetation such as the tower and the sky. Removing such information leads to a more accurate analysis.

There are two types of masks integrated with our approach: community masks (IDE; OGUMA, 2010) and species masks (NAGAI et al., 2011). The former is used when the whole community of plants need to be analyzed together in the visualization; while the latter is used when a particular individual of a species is of interest. With the aid of biologists, we were able to create multiple masks that fit one of these categories. Applying a different mask is an interactive procedure allowed by the prototype in our approach. Figure 4.2 shows the ROI regions from three species. These were defined by (NAGAI et al., 2011).

4.1.2 Step 2: Data transformation using a vegetation index

Phenological analysis from images needs to consider the chromatic coefficients associate to each pixel in the image. The RGB color model is inadequate for phenological studies because they are susceptible to variations in scene illumination. Although image calibration might help (NAGAI et al., 2011), the traditional way, to do such studies, is to transform each pixel's RGB value into a new vegetation index which guarantee a semantic value. There are four semantically significant vegetation indexes: the green and red averages, and the green and red excesses (see details in Section 2.2). We have implemented all four vegetation indexes, nevertheless PhenoVis defaults to the green average. The PhenoVis default green average vegetation index tells how much green is the pixel when compared to the red and blue channels. We have normalized its possible values

Figure 4.2: Three regions of interest: label A indicates species *Betula ermanii*, label B has *Quercus crispula* and label C shows the sample area of *Acer rufinerve*.



Source: Author.

between zero (no green) and one (full green). These intensity results are directly related to the level of green pigments (e.g. Chlorophyll) of the leafs in the trees' canopies.

The phenological analysis of the images focuses in four important phenophases associated to plant development. The checkpoints are: leaf expansion (when plants start to go green), peak or maturity (when leaf growing reaches a plateau), leaf fall or senescence (when leafs start to fall), and post leaf fall (end of growing). The phenophases and the size of the growing season (from start to the end) are indicators of plant development across different years.

The green and red components are a good indicator of the focused checkpoints. Various vegetation indexes formulated using these coordinates are explained in (SONNENTAG et al., 2012). The *RGB* chromatic coefficient is defined by dividing each component (*R*, *G* or *B*) by the sum of the other components (R+G+B). For instance, the average plot of green chromatic coefficients (g_{cc}) is explained to be a good indicator of the above-mentioned phenophases due to the encoding of green pigments in leaves. In PhenoVis, this and other vegetation indexes (including the cited above) are implemented and used to transform the RGB coordinate of a given pixel into a new value which has its semantic meaning. Some vegetation indexes calculations are described by Equation 2.1 and Equation 2.2.

We introduce a new vegetation index, called *HSV* chromatic coefficient, which is the mapping of the *RGB* chromatic coefficient into the *HSV* color space. In particular, we use the hue component (called h_{cc}). Unlike g_{cc} and g_{ex}, where similar values may be associate with distinct colors, similar h_{cc} values are associated with a distinct color (see Figure 4.3). This index gave us new insights when looking at the data. Color features from image's elements trend to be highlighted when used h_{cc}.

4.1.3 Step 3: Filtering by interval of interest (IOI)

Each vegetation index has values in a particular range. For example, g_{cc} is a normalized between 0 and 1. For vegetation images, the range of g_{cc} found in practice is much

Figure 4.3: Colors with the same g_{cc} of 0.4. On the other hand, the corresponding h_{cc} values are very distinct.

Source: Author.

smaller, usually between 0.31 and 0.50 (ZHANG et al., 2012). Since we compute a histogram of vegetation indexes, we narrow the limits of the histogram to a user-specified interval of interest (IOI). The CPM representation color map focusing on the values in the IOI, being the values out of these limits resumed to the biggest (if bigger) value, or the lowest(if lower) value determined by the IOI.

4.1.4 Step 4: Color palette and histogram granularity

Selecting a good color palette is essential in data analysis (LIN et al., 2013). Colors play a major role, and its first influence takes place in the definition of the histogram granularity in PhenoVis. The granularity tells how large is the interval of values of a given bucket of the percentage distribution. The granularity requires the previously configured IOI and the number of colors in the palette. Such definition creates an interesting trade-off between the size of IOI, granularity and the number of colors available. For example, if the interval is very small and we have many colors, each dimension of the IOI will be semantically irrelevant because of the resulted tiny grain size. The same problem appears if the interval is too large and we have few colors to define the dimensions. The best situation arises when there is a balance between the size of the IOI and the number of colors. The number of colors in the palettes varies between 11 and 30. The typical IOI for the g_{cc} has size of 20%, steering the choice of this number of colors, making each dimension exactly 1%-wide.

A dynamic definition of color palettes is possible in our approach, for convenience we have pre-defined a set of palettes. The palette is shown in Figure 4.1 is the most widely used in our work. This palette has four categories with different colors (brown, green, blue, and purple). Each of the five internal divisions of these four zones has distinct levels of saturation (from dark to light brown, for example). Due to this feature we have zone borders that are made by a very bright color neighboring a very dark color, allowing us to visually separate zones. This classification allows us to identify species' behaviors and visually compare patterns in the CPM as we shown in Section 7.

Since the selected vegetation index influences in the CPM data distribution, preferential palettes are set to some indexes. For instance, when using h_{cc} indexes, the standard palettes are all created using the HSV color system. This rules semantically guarantees that all the green appearance into the CPM refers directly to green areas; all blue appearance refers to snow periods. In others words, it associates the colors involved in the samples to similar colors in the CPM visualization.

4.1.5 Step 5: Calculating the percentage data distribution

A vegetation index can be calculated to each pixel in a given input image. Supposing that each image has a range of 600x600 pixels, we will have 360000 vegetation indexes times days of the year (365). This scenario leads us to a vast amount of data. Instead, what is used in the current articles is: to each image, calculate the average of all indexes, leading to only 365 indexes. Average plots of these indexes help identify phenophases, but this analysis may be misleading since different samples might have the same index. In PhenoVis we store more information about each image than a single vegetation index. We use a histogram to encode the percentage distribution of vegetation indexes in an image.

After defining the granularity of the IOI and selecting the corresponding palette, our approach calculates the data distribution of the values within the IOI according to the size of color palette used. The algorithm is straightforward: for each pixel value in the input image, we find the corresponding dimension in the IOI and accumulate it. For each image, the percentage distribution counts the number of pixels in each interval (e.g. between 0.3100 and 0.3199). We repeat this process for all images of the dataset, obtaining one percentage distribution for each image.

To values outside of the IOI, two solutions are possible: the first is to discard these values since they are outside of the analysis scope. The second approach is to account lower values as if they were in the lowest dimension of the IOI, and higher values as if they were in the highest dimension of the IOI. In our experiments with different phenological data sets, we have found that the second solution is attractive because values outside of the ordinary IOI are very rare.

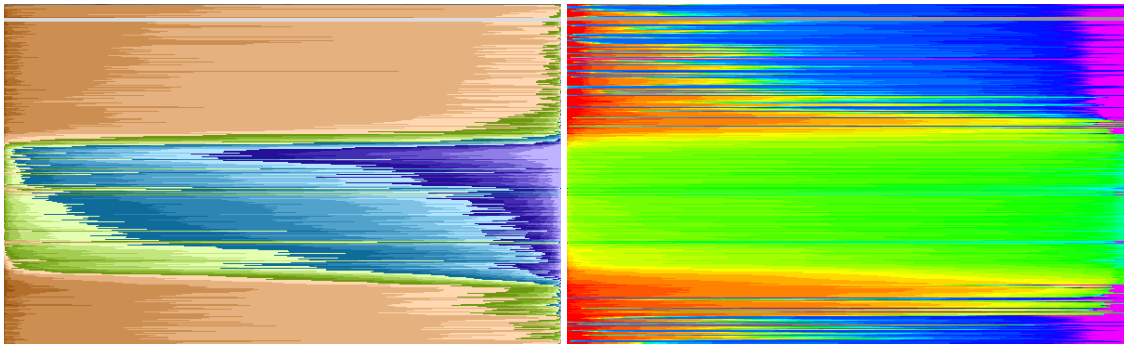
4.1.6 Step 6: Creating the normalized stacked bar chart

To construct a row of the CPM we need to map vegetation index values to colors. A reduction is done in order to fit the mapped pixels into the CPM resolution number. For example, if the image has 6000 pixels and the column resolution of our CPM it is set to 600 pixels, there will be one value for each ten pixels group. This process can be described as “normalization of the frequency of each bucket”. Besides being a normalization, it keeps more information than an average plot. Height and width resolution of the CPM are predefined by the user. The value used in the generated CPMs presented in this work is: 600x366 pixels.

The percentage map of a single input image consists of a normalized stacked bar chart (vertical bar in step six of Figure 4.1). Both landscape or portrait orientations can be used to display the chart. The IOI granularity indicates the number of sections in the chart. In landscape mode, the height of each chart entry is proportional to the frequency count in the percentage distribution of the vegetation indexes. The width is associated with the number of pixels available to draw the percentage map. Colors are defined by the palette being used and their corresponding histogram entries. CPMs generated have 600x366 resolutions, being 366 pixels the number of rows (that represents days in a year), and 600 pixels the number of columns (that represents the resolution of the percentage distribution plot). The map represented in Figure 4.4 (a) shows that sections identified by the light brown and dark green colors are the ones that have the most frequent values for the vegetation index being considered.

The chronological percentage map final chart consists of a sequence of percentage maps stacked in chronological order, from top to bottom (portrait) or left to right (land-

Figure 4.4: CPM examples: (a) categorical color table using the g_{cc} index, (b) HSV color table using the h_{cc} index.



(a) g_{cc} (b) h_{cc}

Source: Author.

scape). In our analysis, each CPM corresponds to all images of a single year. Figure 4.4 illustrates two examples CPM charts for the same year using different vegetation indexes. The first one, shown in (a), uses g_{cc} and the categorical color table. In this example, shades of green, blue and purple respectively correspond to g_{cc} values between $35-40\%$, $40-45\%$, and $45-50\%$. Higher g_{cc} values are associated with greener regions, and a clear pattern of growing season emerges in the middle of the year (these images are from Japan in the north hemisphere). The second example uses h_{cc} and the HSV color table. The growing season is associated with shades of green. This CPM allows us to investigate unusual patterns outside the growing season, which displays shades of red and blue.

4.2 Dataset size importance

The Chronological Percentage Maps are built using the percentage histograms of a set of images collected over time. This representation can detect significant phenological patterns that show up throughout the seasons of one or many years of data. The time constraint is, therefore, critical to understanding broad changes in plant species behavior. Phenological patterns might arise in small and large temporal scale. For example, small temporal scale patterns appear when plants change behavior rapidly within a small time frame such as during the day, weeks, or months. In conclusion, significant temporal scale patterns can only be detected when considering several years of data and reflect broad movements in a plant growth.

4.3 Considerations: Multidimensional increasing discussion

A CPM is generated based on one metric technique. Because of it, many results values from the same method are considerable a unidimensional data. However, if we divide the possible result values of this data in gaps of values (e.g. 33% to 34%), we enable a multidimensional approach. For instance, if we deal with different amounts of pixels distributed by n predefined gaps of values, our unidimensional approach generate a n dimensional approach.

Dimensionality refinement (percentage maps) gives more quality to a mathematical statistics search (such as MAE, MSE, MAPE, and Kullback-Leibler). As consequence, this detailed information feature adds a finer characteristic to our approach if compared

to the average plot. Representation matrix with 100, 1000 and 10000 buckets were implemented and tested in our prototype. For instance, the values that were resumed to one bucket in a matrix with 100 dimensions, is spread in 10 buckets to a matrix with 1000 buckets, and so on. Once the matrix precision increases, more significant results can be seen as well.

5 PHENOVIS – DATA ANALYSIS

There are a number of questions that drive plant phenological analysis, such as the comparative analysis of data from different years. For example, a comparison of the starting dates of a given phenopase (e.g. start of leaf growing) allows to verify how these values changed with the progression of the years. Automated similarity comparison of different phenophases is important in this process.

PhenoVis leverages the fact that more data is encoded in percentage distributions to provide a more sophisticated automatic similarity analysis. The user can select a specific time interval directly over the CPM of one year, and PhenoVis suggests a similarity rank of the other years based on this pattern. PhenoVis is also capable of detecting similar years without any user interaction, presenting search results graphically. The data mining feature coexists with the CPM, using the same information that is calculated to create them.

This section details the design of PhenoVis regarding automatic search algorithms. Next subsections present the basic setup, the algorithm for ranking years by similarity, and the visualization feedback to support the analysis.

5.1 Similarity metrics

In PhenoVis, we have adopted four existing and well-known similarity techniques for the automatic search: the Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), and the Kullback-Leibler Divergence (KLD). Each of these mathematical techniques presents characteristics that might change the results significance. The MAE is a usual estimator of the difference between two matrices. Used for the same reasons, MSE highlights significant gaps between values because it each the distance is calculated by power of two. MSE is a very risk technique when it is needed to deal with noisy dataset. 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 two probability distributions.

All techniques operate uniformly, taking as input two CPM subsets for the comparison. The first parameter is the CPM query, defined by its start and end dates. The second parameter is the CPM candidate, which requires the same number of days of the query, but may differ from start date. 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 a final similarity error. The closer this value is to zero; the more similar are the CPMs. Being P a vector of n predictions, and Q the vector of the true values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - Q_i| \tag{5.1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - Q_i)^2 \tag{5.2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|P_i - Q_i|}{Q_i} \tag{5.3}$$

$$KLD(P||Q) = \sum_{i}^{n} P(i)\ln\frac{P(i)}{Q(i)} \tag{5.4}$$

5.2 Search window

The query pattern defines a time window 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, called fixed window, looks for CPM subsets with matching start and end dates. This comparison returns the years that have the same pattern happening in the same days of the year. In the second way, called moving window, the search looks for the same pattern but without a fixed starting day, allowing the window to move along the year. This search allows to find patterns in other years that happened at different time of the year (e.g. a late growing season).

Fixed Window For this purpose, a search window $w_i(d_{min}, d_{max})$ is defined by the first and last day respectively (d_{min} and d_{max}) of the CPM of a given year y_i. The search looks for the same window in the CPMs of the remaining years ($y_j, j \neq i$), and returns the closest years using a similarity metric chosen by the analyst. For example, if the analyst selects a window of 2007 that goes from day 56 to day 100 of that year, the pattern search considers as candidates for the comparison only the time window 56–100 of the remaining years.

Moving Window In this case, the size of the search window is fixed, but the search looks for the closest window of the same size of the search window, but allowing it to start at any possible starting day. This search algorithm is more expensive but gives better results than the previous since it allows pattern drifts to happen, which is relatively common and even expected in phenological data. Following the previous example for the query pattern from 56–100 days of 2007, this algorithm might find, for example, that the best match according to the similarity metric is of year 2010 from days 65–109. To help the biologists when using the moving window search, we output an additional attribute stating if the best match happened earlier or not considering the starting date of the query pattern. In the example given, this attribute has the value of nine, because the best match in 2010 appeared nine days later in that year.

In the case of the fixed window search, the similarity metric is calculated only x times, where x is the number of years available minus one (the year used to configure the query pattern). Whereas, the moving window, has the similarity metric calculated for all the possible combinations of the query pattern and the moving window pattern. Therefore, the search iterates as the equation presented below. The attribute *yDays* is the number of days in a year(365). The number of years in the dataset is represented by *dyNumber*(TKY: 10 years). The *tGap* is the temporal gap size (in days) from the query.

$$SearchedTimes = (yDays - tGap) * dyNumber; \tag{5.5}$$

5.2.1 Search using filters

We also implemented a filter that allows only a subset of the percentage distribution to be taken into account. The target pattern is a temporal slice of a chronological percentage map of a given year. The pattern can contain all the channels values of the interval of interest or only a subset of them. Restricting the pattern configuration for a subset of the IOI may enable automatic search only for relevant dimensions from the analysis point of view. Once set, the target pattern is searched in the remaining available CPMs. Consider the case that the expert is interested only in g_{cc} values in the interval of 40% and 45%. The similarity search can be set to filter out values outside this interval. Even using these filters, the search still can be performed using the two different proposed ways: fixed and moving window .

In phenological datasets, the absence of data (days where the collection system malfunctions) and the presence of outliers (missing data or odd days) are very common. These situations are problematic for the automatic search. We also have an outlier filter that removes outliers from the similarity search. An automatic outlier detection was implemented taking into account the percentage distribution of a given day and its adjacent days distribution. The problem is tackled by automatically detecting outliers by comparing each percentage map with the surrounding percentage maps using a sensitivity parameter defined by the analyst. If the percentage map is too different from the neighboring days, we consider it an outlier. Outliers are finally kept out of the calculus of the similarity metric, and the respective percentage maps of the candidate for comparison is also ignored when the similarity computation occurs.

5.3 Ranking visualizations

PhenoVis is also capable of ranking years by similarity according to the target pattern set by the analyst and the selected similarity technique. Instead of showing only the best match considering the query pattern (as shown in Section 5.1), we list a rank from all the years available in the data set. The error value calculated by the similarity algorithm defines the rank. Even if the matches are slightly different, the ranking gives an idea of how the variation patterns appear along time.

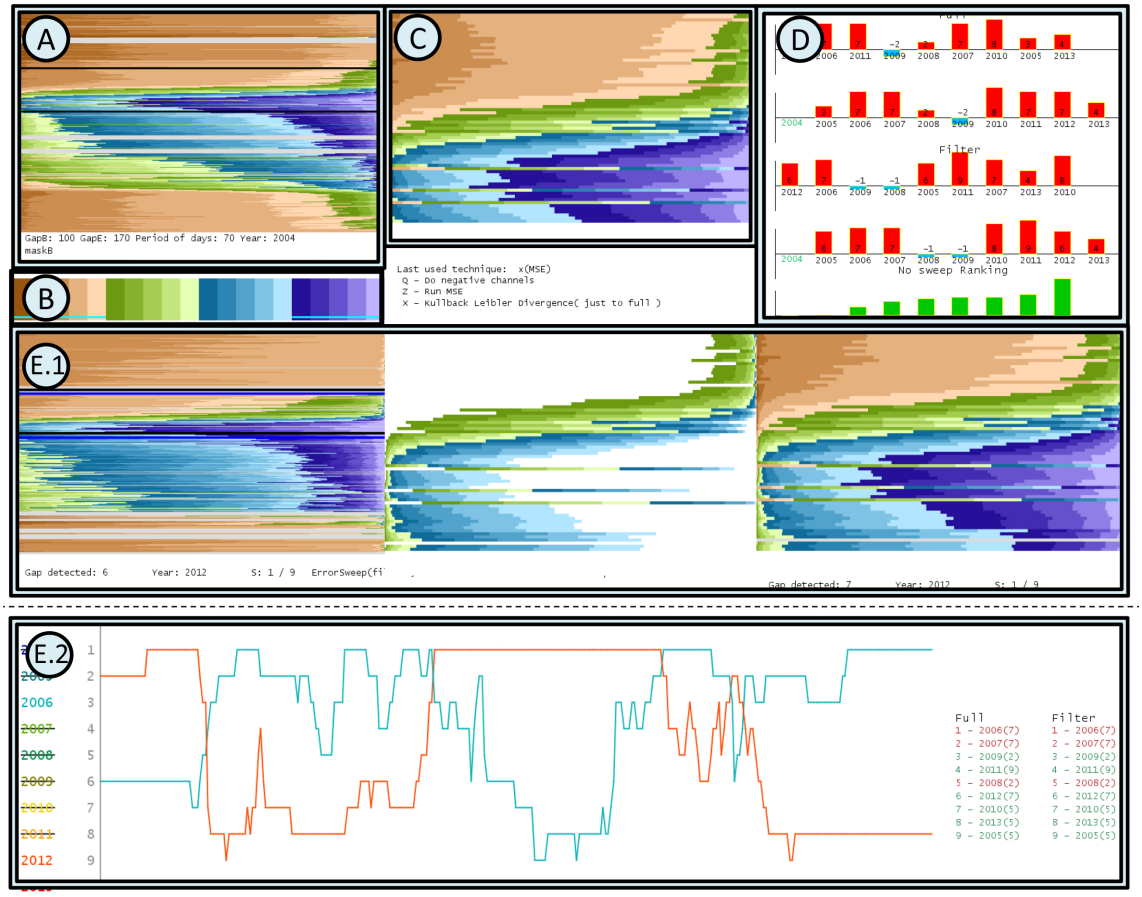
When a fixed window is used, a single search is made to each year presented in the dataset. It results in an error for each year comparison. All errors are normalized and displayed as a bar chart graph in order to allow comparison between them. When a moving window is used, the difference in days between the query event and the most similar found activity need to be represented as well. Another normalized bar chart is used to display this information.

Two ranking visualization modes were created in our prototype in order to compare more than two temporal datasets at once. The first is the single-rank visualization mode. This was created to present more details about the similarity search results and allow a finer search control. The second mode is the general-rank visualization mode. Since this mode aims to represent a wider range of results, it was designed to simplify information. On the other hand, it allows the single-rank visualization mode access by just clicking on any desired sub-block. Both modes support moving and fixed window search.

5.3.1 Single-rank visualization

This mode displays results ordered by the distance computed using the similarity metric; a single time gap (event) is compared with others years samples. Figure 5.1 describes the single-rank elements. This ranking main goal is a more detailed display of the results. It also allows new similarity searches based on the last results. In addition, the searches in this mode are finer once it allows channels selection.

Figure 5.1: Single-rank Visualization. Area (E) has two possible displays (depends on the interaction). (E.1) shows details from the similarity search results, one-by-on. In case of a moving window search, a rank variation of the years (y axis) against the start date of the query (E.2).



Source: Author.

To setup a more detailed search, some interactions can be done on the results in order to change some attribute and keep others. The query's CPM is displayed by Figure 5.1 (A). Set a new start and end date of the events (by moving the black horizontal line) is also possible in this area. The color palette is displayed in Figure 5.1 (B). Enabling and disabling channels in the color palette allow the filter search by channels.

Query events are zoom in Figure 5.1 (C). The zoom intensity value is scaled in order to fit on the screen. Days distance graph is a bar chart that inform how anticipated, or postponed the similar resulted events were from the query event. These graphs are sorted by error value and year, and are displayed in Figure 5.1 (D). The bar size and orientation (up or down) represents the time shift among the events (query and founded). The green graph plot is sorted by the error value. The bar size here is the normalized error, where the

most similar result is always normalized to 0, and the most different result is normalized to 1. Each of the results and similarly founded gap can be seen individually using a CPM representation, see Figure 5.1 (E.1). A line graph, which allows the analysis of the ranking variation among the sweep of the moving window method, is shown in Figure 5.1 (E.2). These features allow a more specific analysis of the event.

A study case is presented in this paragraph in order to aid the understanding of the rank variation from Figure 5.1 (E.2). The year of 2012 (red line) first queries results in the first positions of the rank fowling by a huge steep when the winter ends. After that, 2012 oscillates its positions a bit. However, it stays in a lower positions of the rank until the next pick, in April. With the summer beginning, the samples comparison starts to get similar results. It increases the graph to the first position of the rank where it remains stable until the end of the season. Subsequently it gradually falls to a lower positions as the cold seasons starts. It remains there until the end of the year.

5.3.2 General-rank visualization

This mode's motivation is to present a wider range of results in a high-level way. It also allows to compare an all years set of samples from a mask *A* against all years samples from a mask *B* (different masks), as well as identical masks.

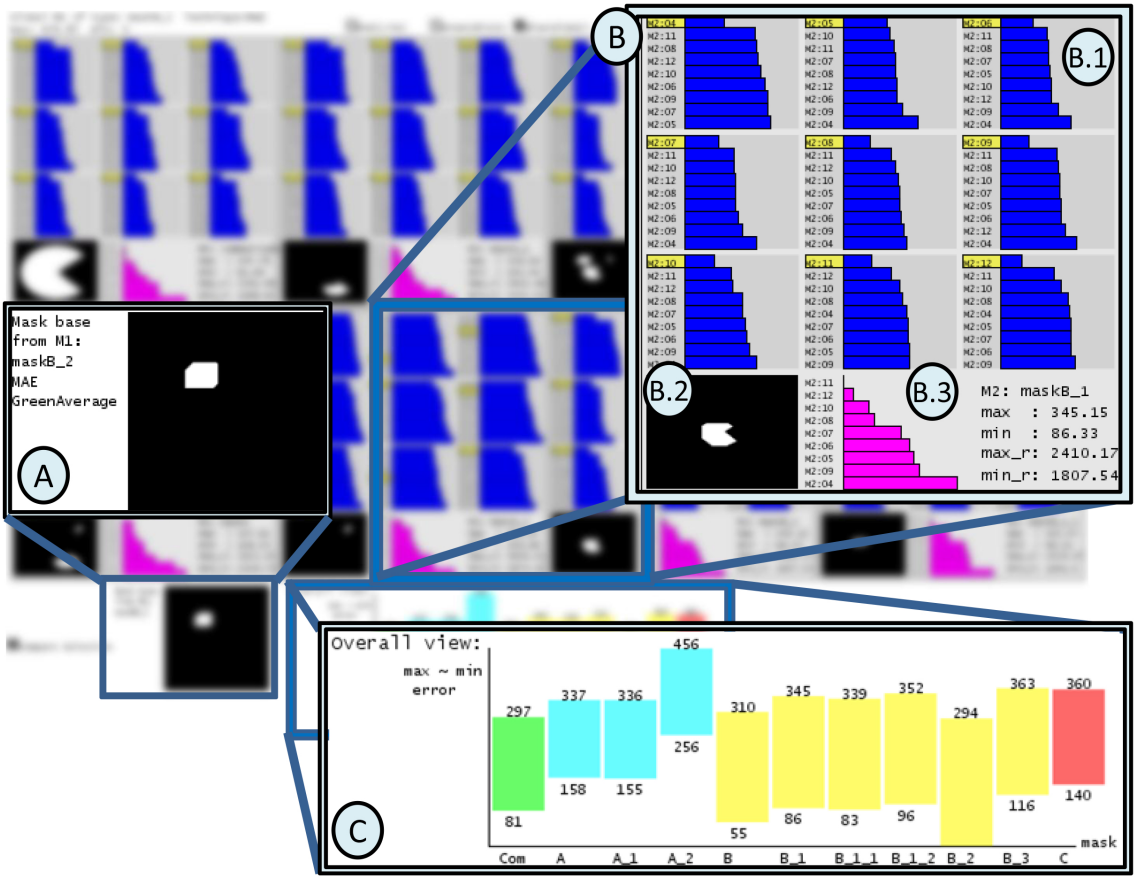
The general-rank visualization is resulted from a set of searches that involves multiples combinations of results comparisons. As consequence, it requires a set of configurations. The technique that is going to be used by the search (MAE, MSE, MAPE or KLD), a central mask to be compared, and a set of other masks that we want to compare it. The time gap of the CPMs that we will be using to search similarities. Pre-set time gaps that were specified in (NAGAI et al., 2011) can be used: Bud dormancy, leaf expansion, peak, leaf fall and post leaf fall. It is possible to change the order that we are going to display the year's results into each comparison. The years can be sorted by metric result value or sorted by the temporal order. The last parameter determines the moving window or a fixed window approach to the search.

The mode is divided by blocks of results. Each represents a compared mask, and are composed by sub-blocks that display the errors generated by the years comparisons in bar charts. For instance, to a range of years starting in 2004 and ending in 2012 (9 years), each block will contain nine sub-blocks with nine error's bar chart each. Another bar chart called general rank is displayed at the bottom of this mode to overview the maximum and minimum error generated by each mask comparison.

Each sub-block has the same range of content of a single-rank visualization. Even being less detailed it allows the access to the single-rank visualization by clicking in one sub-block. It is also possible to scroll among the results if the number of blocks ends bigger than the screen support.

All labels reference in this paragraph are from Figure 5.2. Each block (B) refers to a years comparisons of all year's samples from mask (A) against all years from (B.2). The highlighted years (yellow) of each sub-block the respective query year. Element (B.1) is divided by 9 sub-blocks(blue charts). Each refers to a single year comparison from a mask (A) against all years from (B.2). The block general rank is generated by the normalization of the sum all error results from each year. It aims to turn visible the relationship between the years. The maximum and minimum error value from each block is displayed in (C). Each chart represents a block. The colors are sort by species mask categories. Elements, that we also display inside each block, are: the status of the years and errors value limits.

Figure 5.2: General-rank Visualization. (A) displays the mask from the query matrices, in this case $maskB_2$. (B) structure represents a block. (B.1) is divided by 9 sub-blocks(blue charts). (B.3) is a general rank from the block. The overall view (C) aggregates information about each block.



Source: Author.

6 PROTOTYPE IMPLEMENTATION

The Chronological Percentage Maps have been implemented in a prototype written with a Java-based programming language called *Processing*. As an input, the prototype receives a set of raw images grouped by year and an image mask that defines the region of interest of the analysis. The prototype works in two phases: images pre-processing stage and PhenoVis interaction phase. However the images pre-processing stage is executed only once to each dataset, being the PhenoVis interaction phase the most usual one.

Pre-processing stage applies the chosen mask in all the raw images, generating the CPMs matrices that are used in the analysis. This phase is a computationally expensive operation. To generate all CPM matrices files for a dataset of 10 years (3449 images samples) takes around 6 minutes on a modern computer. However, this latency mainly occurs because of the image's load effort, and not because of the matrices generation algorithms. The prototype automatically saves the generated CPMs matrices in permanent storage in order to execute the pre-processing only once. Since the matrices are made, the raw images are no longer needed. In the PhenoVis interaction phase, the prototype uses the percentage maps (CPMs matrices) created in the previous step in order to offers the analyst a graphical interface where the CPMs can be interactively manage and analyzed.

Since matrices are generated by PhenoVis in order to create CPMs, execute patterns find, and ranking analysis. The single set up of the set o matrices is the grain size of it. As more detailed a matrix is (more grains), finer is the search. In addition, the needed space in disk increases proportionally with the detail characteristic. Our prototype supports matrices with the grain size of 100, 1000, and 10000. The table 6.1 exemplifies the matrix relationship among the grain size, generated file size, generation time, fixed window search (FWS), and moving window search (MWS).

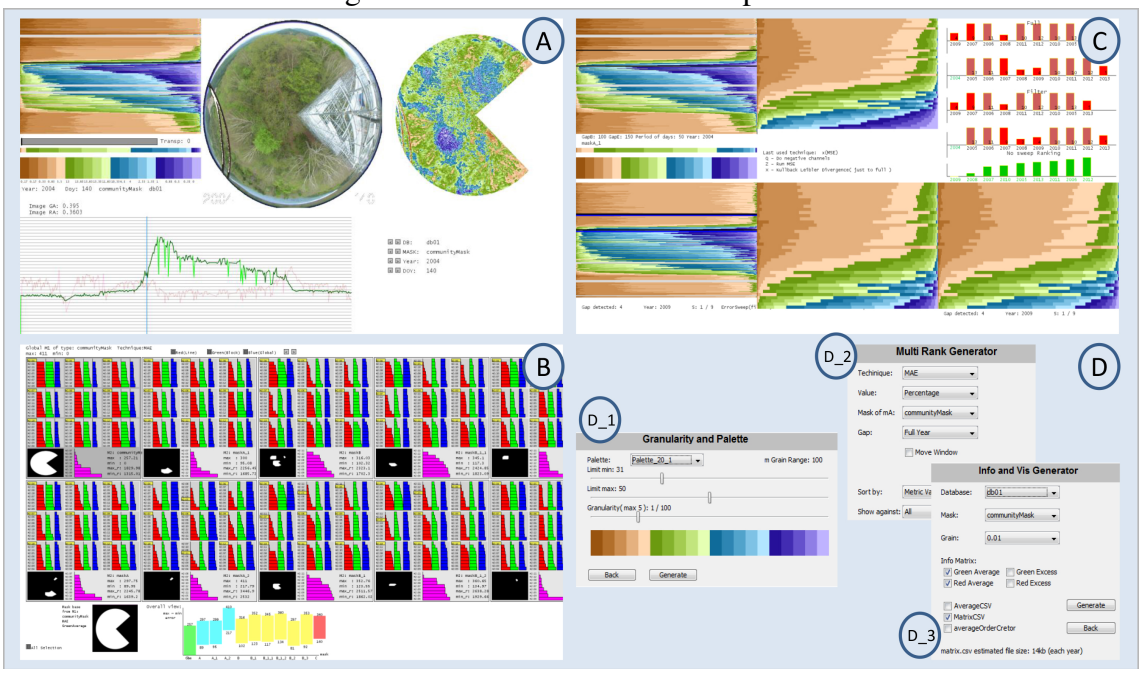
Table 6.1: Matrix features increasing in relation to the grain size.

Grain size	File size	Creation time	FWS	MWS
100	82 Kb	11 seconds	$6 * 10^{-9}$ ms	$1968 * 10^{-9}$ ms
1000	824 Kb	12 seconds	$7 * 10^{-8}$ ms	$2461 * 10^{-8}$ ms
10000	8.2 Mb	14 seconds	$9 * 10^{-7}$ ms	$2947 * 10^{-7}$ ms

Execution time from searches (FWS and MWS) and file size increases linearly in relation to the grain size. Matrix creation time is the sum of 3 process. It involves load each image, extract its RGB features and calculate vegetation indexes for each pixel. A grain increase does not imply in load more images; thus the time involved in this first part (load images) of the process is the same to any grain size configuration. However, the logic process of calculating vegetation index has a linear time spending increase in relation to the grain size. Since load each image from disk to memory and extract its information is the most expensive part of the matrix generation, the results presented above have not a linear increasing.

Figure 6.1 shows the three primary interaction modes available in our tool: Figure 6.1 (a) similarity search mode; Figure 6.1 (b) general-rank visualization mode; and Figure 6.1 (c) single-rank visualization mode. These exploration modes allow interact with the results, and visual analyse the generated CPMs. Control pannels Figure 6.1 (d) allow the

Figure 6.1: PhenoVis interaction phase.



Source: Author.

configure features such as: color palette change, IOI and granularity set Figure 6.1 (D_1), edit search variables to generate a new Multi-Rank Visualization Figure 6.1 (D_2) and aid the addition of new datasets Figure 6.1 (D_3). Since most of the data required by the second phase (data exploration) is already processed, the application is quite responsive even when the MAE, MSE, MAPE and the Kullback-Leibler algorithms are executed on scenarios with several years of data.

6.1 Data Acquisition and Data Standardization

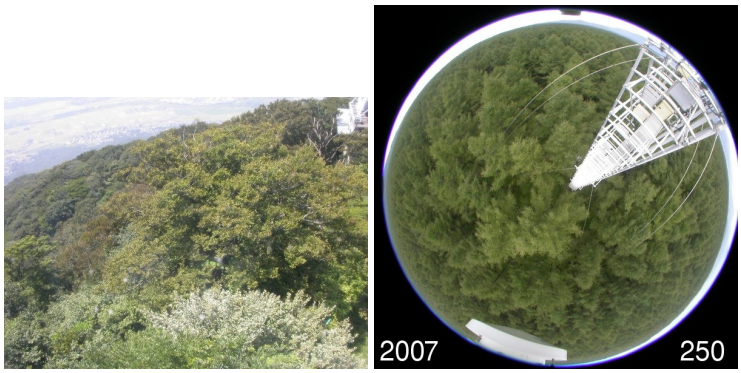
The five datasets used in this experiments were gathered by the Phenological Eyes Network (PEN) (S et al., 2005; NISHIDA, 2007) and are publicly available on their internet site (http://www.pheno-eye.org). Usually, phenological data consists of several digital photographs taken by static cameras set in remote monitoring forest regions. Cameras placed in tall towers; are equipped with fish-eye lenses to capture a larger area of vegetation. Fish-eye lenses have a shorter focal length than the diagonal of the projected image. As a consequence, a wider field of view is obtained giving much more coverage of the forest under study.

We report results for the most interesting datasets that we found, which have data from multiple years since 2004 (variates to each dataset). These datasets are listed bellow:

- Takayama Flux Site (TKY), located in Japan, and has a deciduous broadleaf forest. We use regions of interest to identify three species: *Betula Ermanii* (Species A), *Quercus Crispula* (B), and *Acer Rufinerve* (C).
- Alice Holt (AHS), located in Hampshire, UK. This is a commercial managed region from lowland oak forest. The main tree species is Quercus robur L., but other species, including European ash (Fraxinus excelsior L.), Q. petraea (Mattuschka) Liebl. and Q. cerris L., are present.
- TERC Grass Field (TGF) content is a grassland data managed by U. Tsukuba and AIST.
- Mt. Tsukuba (MTK) is a mixed forest dataset from the vegetation presented in the Mount Tsukuba Japan. MTK dataset was not collect from a fish-eye camera. Nevertheless, we analyse the vegetation area equally, ignoring the sky in the same way we ignore towers in others datasets.
- Fuji-Hokuroku (FHK) , located in larch forest ("deciduous needleleaf forest") in the foothills of Mount Fuji, Japan.

Figure 6.2: Image samples from datasets used on PhenoVis study and development.

(a) TKY (b) AHS (c) TGF



(d) MTK (e) FHK

Source: Author.

Besides using all the listed datasets to evaluate CPMs generation and temporal patterns search, we analyzed the dataset from Takayama. As long as we had information about specie's canopy locations from this dataset (provided by (NAGAI et al., 2011)), we were able to add species' analysis in our work. The images used to create the CPMs are standardized on a single representative image of each day with 600 by 600 pixels. These images are in the JPEG file format that, for each pixel of the picture, contains an intensity value of red, green, and blue. These values are used as input in PhenoVis.

7 VISUALIZATION RESULTS AND ANALYSIS

In this Chapter, we present a visual evaluation of the CPM. We compare it with the average-based plot, discuss color palette interference in the analysis, and present the species' signatures feature. A section of evaluation of the proposed general-year data analysis algorithms and display modes is divided three parts. First, we present the results from the similar phenological patterns searching. Next, comparisons among the phenological patterns results from different species are shown and commented. The section ends with the visual analysis of the ranking visualizations.

7.1 CPM evaluation

We present here results related to the CPM visualization: a comparison with the traditional average-based plots of phenological studies, CPMs expressive power, and the using of the CPM as a species' signature.

7.1.1 Comparison against average-based plots

Two examples are shown by Figures 7.1 and Figure 7.11(displayed and commented in 7.3.3) in order to demonstrate differences among CPMs and average plots. CPMs' detail sensibility and visual flexibility are shown next, as well as a subsection that link the use of h_{cc} with its main features.

Figure 7.1: Two groups of four percentage maps with the same g_{cc} average: they look different using PhenoVis' CPM representation. The first Group refers to samples presented in Figure 1.2.

(a) $g_{cc} = 0.3905$

(b) $g_{cc} = 0.4442$

Source: Author.

Figure 7.1 illustrates the difference between CPMs and averaged data. It shows two groups of four percentage maps (based on different days of the TKY dataset) with the same $g_{cc} = 0.3905$, and the same $g_{cc} = 0.4442$ respectively. Although they have the same g_{cc} average, they appear differently in the CPM representation. Such differences might indicate unusual phenological patterns. CPMs can differentiate between images considered equivalent by a given vegetation index. Besides the lack of results from vegetation

indexes such as b_{cc}, r_{ex}, g_{ex}, and b_{ex}, these are implemented in our prototype and are ready to be explored by the experts.

7.1.2 Palette interference analysis

Mapping values to colors are one of the main steps in the CPM construction. Since each palette has a feature, to choose the best one to support the desired solution can be a hard task. We have selected two of the available palettes in our prototype in order to highlight different types of features from each one. Figure 7.2 shows a mapping results to both palettes using g_{cc} and h_{cc} as vegetation indexes. To g_{cc} translations, we used IOI between 0.31 and 0.5. On the other hand, since h_{cc} has a different technique involved in translate values to colors, it uses a full range IOI.

In order to highlight features from Palette 1 and 2 (from Figure 7.2), Figure 7.4(a) shows four different samples with the same average of green intensities (see also Figure 1.2). Using the zone color palette, we can observe how the percentage maps look different from them. Just by looking at percentage maps of 2007 and 2012 pictures, we can see that the brown shades channels are much more present in 2012 than in 2007. This same effect also appears considering the whole chronological percentage map, of a full year.

Segmented areas detection becomes worst when Palette_1 (color zones) are used as input images. However, they appear to be much better choice when used in percentage maps, as shown by the comparisons depicted from (b) to (d). For the year of 2004 picture, Palette_1 gives a better result when applied to generate percentage maps. Zones represented by the green shades channels are the most present. In this case, such different channels with the biggest amount of the percentage distribution can be used as identifiers in future comparisons. These are also suitable candidates to suffer a finer analysis through using tools such as zooms and channels filters.

7.1.3 CPM as species signature

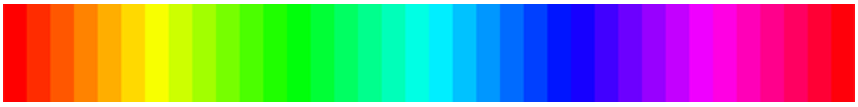
The input images in the forest of the TKY dataset have multiple plant species, and often phenologists need to perform a species' identification. While most of this process is done manually by specialists, an automatic identification is desired due to the large amount of data to process. In this subsection, we evaluate if the expressive power of CPMs can be used as a species' signature. For this analysis, we used the species' identification presented in (NAGAI et al., 2011) for the TKY dataset. In order to represent each species' region of interest correctly and allow a bigger range of comparisons, we created separated CPMs. These were divided by three types: composed by a single canopy, multiple canopies individuals from the same specie, and community CPMs (that includes all known individuals).

A single percentage map corresponds to a sequential align bar chart. The placement of adjacent percentage maps generates visual cues along the time axis. By exploiting this characteristic, we can see that different species creates different CPM patterns. Figure 7.4 shows the CPM for the A, B and C species using the g_{cc} vegetation index for 2007 and 2008. The three most relevant index intervals are green ($35-40\%$), blue ($40-45\%$), and purple ($45-50\%$). These are very similar intra-species for the two years. However, they have distinct patterns when comparing one species against the other, especially on leaf growing and senescence phases. Species B has a purple zone that shows a peak from the greening phase, which stabilizes in the rest of the growing season. In this analysis, the use of g_{cc}, and the categorical color table was helpful to identify different patterns.

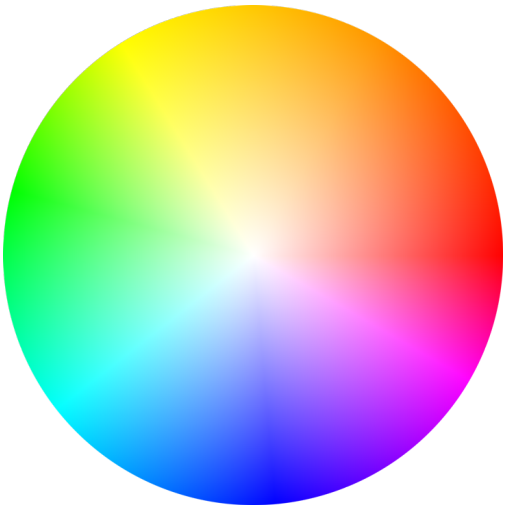
Figure 7.2: Three color palettes that were used in our approach (a, b ,and c). Example of translation using g_{cc} and Palette_1 from (d) to (e) and using Palette_2 from (d) to (f). To generate (g) from (d), h_{cc} index was used with Palette_3, which was created based on HSV cylinder.

(a) Palette 1

(b) Palette 2

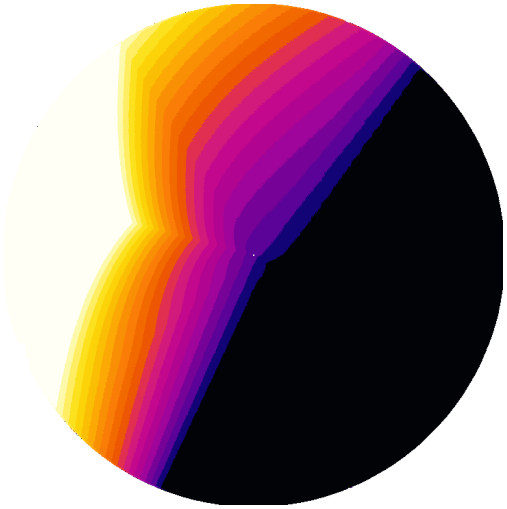


(c) Palette 3

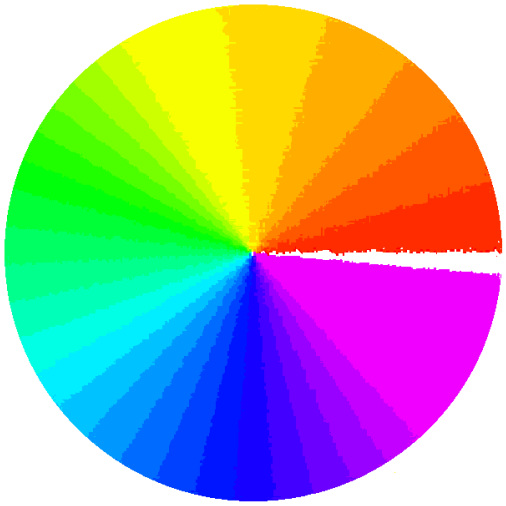


(d) RGB image

(e) Palette 1 - g_{cc} - Translation



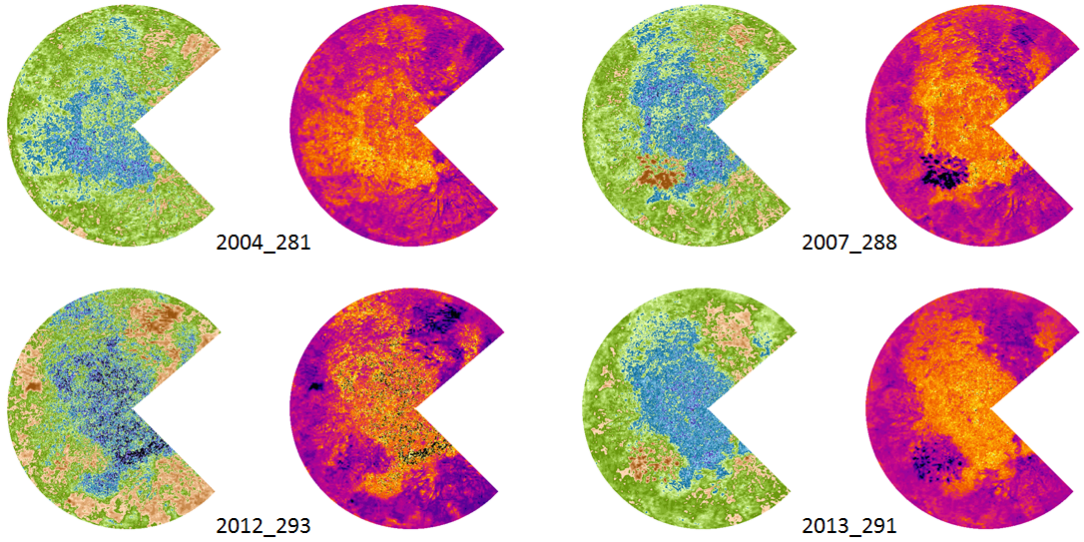
(f) Palette 2 - g_{cc} - Translation



(g) Palette 3 - h_{cc} - Translation

Source: Author.

Figure 7.3: Distinct images with the same average are used to illustrate our choices for color mappings. While the sequential colormap is more suitable for identifying patterns in the input image, the zone colormap provides more intuition on the distribution within the CPM.



(a) Same average images translations

(b) 2004

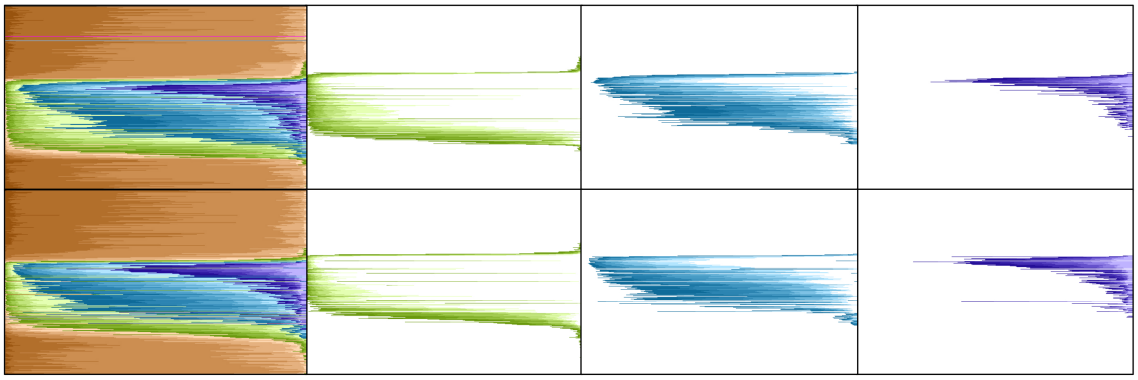
(c) 2007

(d) 2012

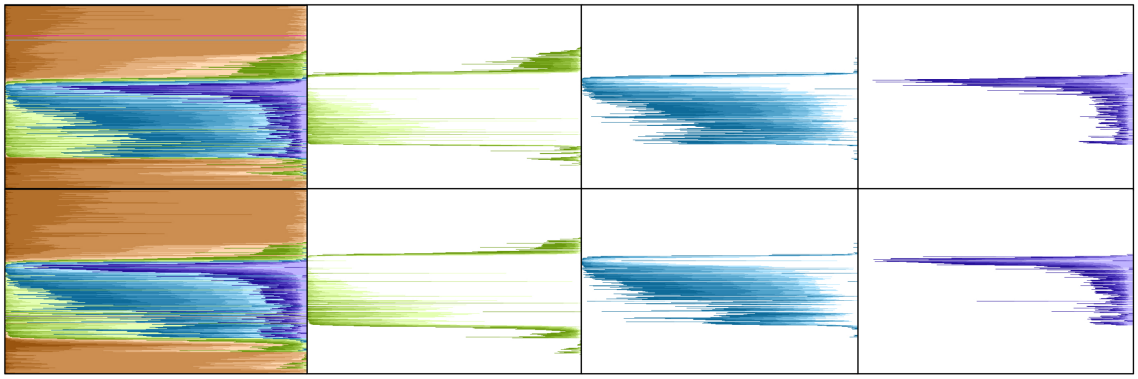
(e) 2013

Source: Author.

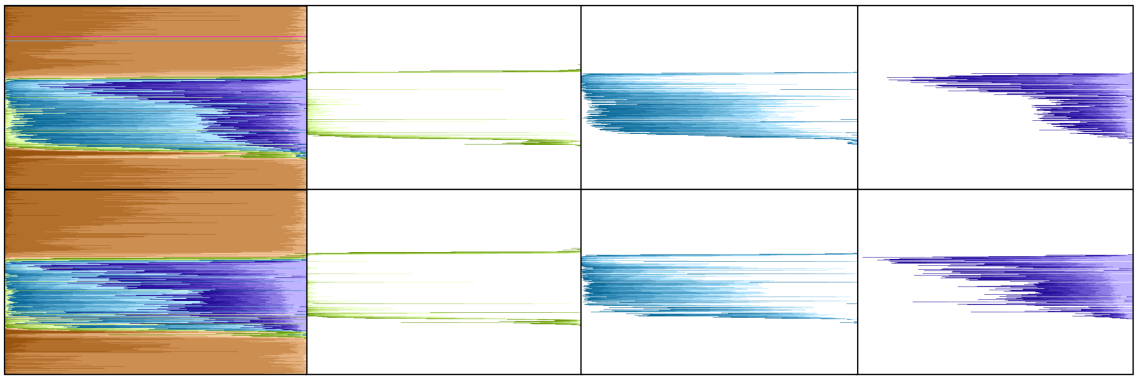
Figure 7.4: Time is displayed top to bottom, being the first item the first day of the year. CPMs and zone highlights for three species for years 2007 (top row) and 2008 (bottom). Observe the distinct patterns in the three zones (green, blue, and purple), which can serve as a signature to identify a given species.



(a) Species A



(b) Species B



(c) Species C

Source: Author.

The three most changeable zones (green, blue, and purple) are very similar over the years to samples of the same species. It is also possible to identify characteristics of the data. Temporal gap representation of growing seasonal phenomena, especially on greening and senescence phases, shows very particular behaviors of each species' individual. For instance, the main zones associated with the *Quercus crispula* sample showed in Figure 7.4 (b) are:

Green zone It has a clear green zone precipitation during the greening phase and the delay behavior during the senescence phase.

Blue zone Even being visually very similar to the other species samples, it can be seen that for the species of Figure 7.4 (b), it presents most part of the samples during the growing phase.

Purple zone After the peak from the greening phase, this zone is stabilized throughout the rest of the growing seasonal. It also disappears suddenly, different from the behavior of other species that follow an order of the gradual disappearance. There is clear a green zone decrease as it can be observed in the purple zone stability after the greening phase.

One of the explanations of these patterns from the biologists is that our technique provides a view from how the query event influences each species. It happens because the CPMs are reflecting phenological phenomena, depicting these smooth transitions between a phase and the next. The relationship of the line neighbors determines the geometric shape of a CPM. The neighborhood similarity or distinction directly influences a characteristic of the final geometric shape. The values variation of each percentage map (channels) over the time, are directly related to the final size of these forms. If we analyze the CPM as a group of curves, the positions of the colored representation in the line is also directly related to the curve derived trend. Analyzing the neighborhood behavior allows us to identify species patterns as well as identify outliers.

7.2 General-year data analysis

We present here results related to PhenoVis' data analysis capabilities and search for similar phenological patterns.

7.2.1 Searching for similar phenological patterns

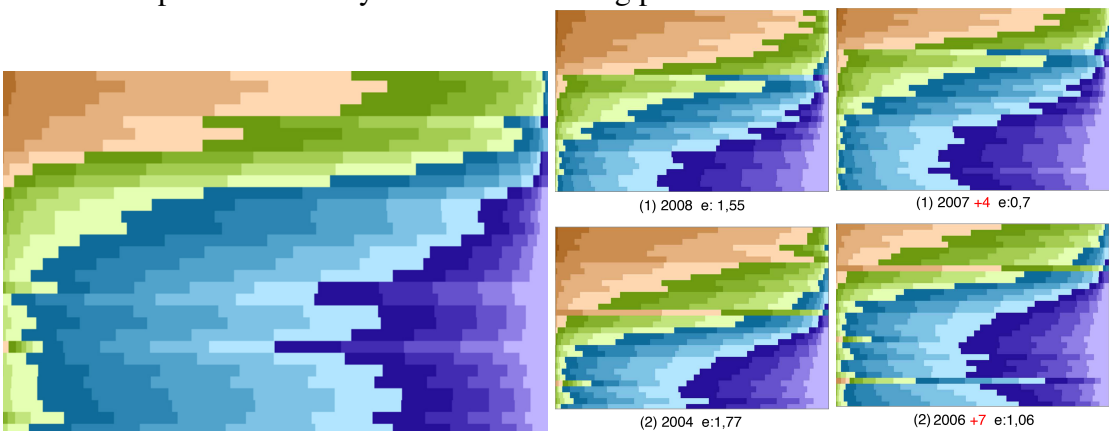
PhenoVis allows the user to specify a window of time over the CPM to define an interesting pattern to be searched. This model will be used as a query for similar patterns in other years. Figure 7.5 shows the results using as query window the leaf expansion period of 2009, and the MSE similarity metric. The results obtained show that this pattern was more similar in the years 2008 and 2004 for a fixed window. On the other hand, it was closer to 2007 and 2006 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. The best two results for a fixed window search show that CPM patterns have noticeable differences. Moreover, as expected, the moving window approach presents a smaller error than the fixed window.

The experts can better characterize phenological behaviors by selecting only a part of the interval of interest of the vegetation index. For example, using g_{cc}, the biologists could conclude that the Leaf-Expansion is mainly defined by a climb of the values 33%, 34%, 35%, and 40%. Thus, a search for these particular occurrences (ignoring others values) can be done (see Figure 7.6). This feature enabled sophisticated phenological insight to the biologists. In addition, we have observed that the KLD similarity search works better in such scenarios. On the other hand, MSE and MAE are best options for a search that considers the full IOI.

7.2.2 Year Ranking

In a single-rank analysis, the user selects a query pattern of a given year. We can then rank the appearance of that particular pattern over the remaining years of the dataset,

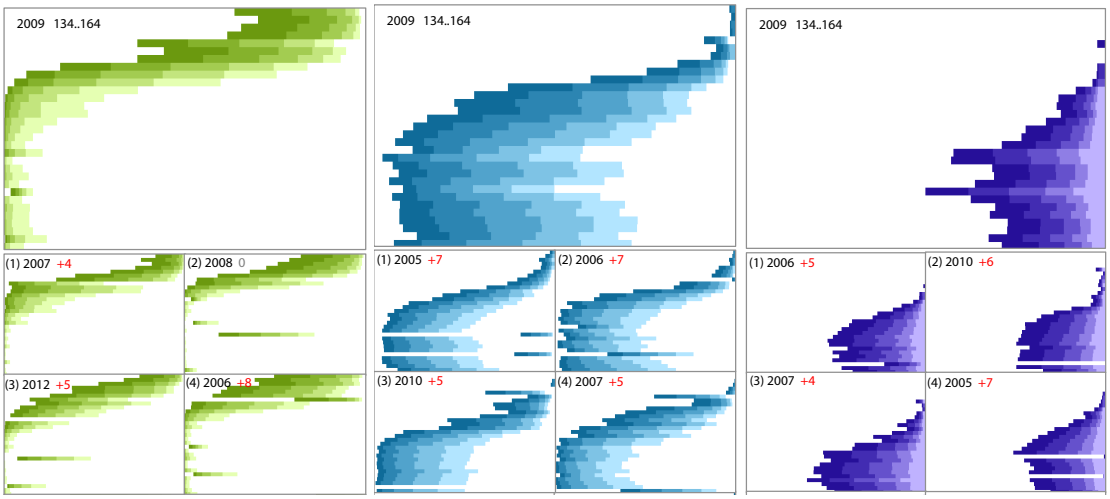
Figure 7.5: MSE similarity search: (a) query pattern; and top two results using (b) fixed and (c) moving window. Comparison errors are given after the letter "e"; red numbers show the temporal shift in days from the matching pattern.



(a) Query pattern (days 134 to 164 of 2009) (b) Fixed window (c) Moving window

Source: Author.

Figure 7.6: Finding similar regions filtering by specific zones using KLD.



(a) 36-40% zone (b) 41-45% zone (c) 46-50% zone

Source: Author.

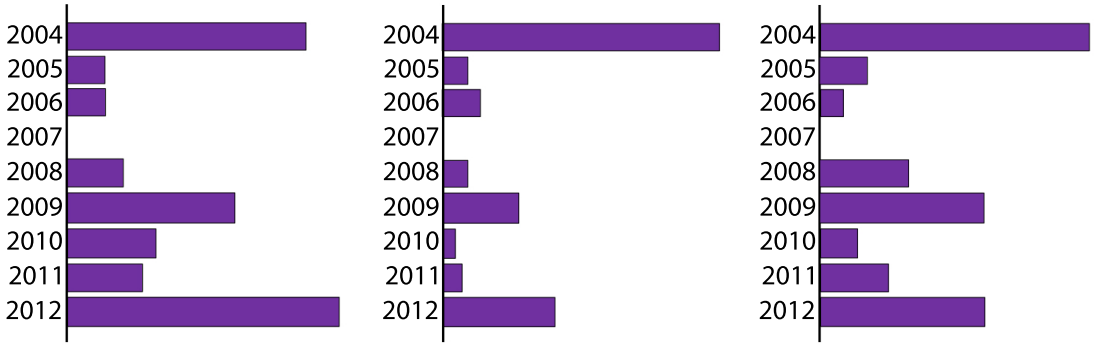
using one of the similarity techniques. The general-rank does the same similarity check but considers all the combinations of years according to the selected temporal window of the pattern. As result, we observe homogeneous groups of years, ranking how much similar is a given year in respect to others.

Some patterns and trends can be seen in the TKY dataset by using our approach. Figure 7.7 shows the results obtained with a year ranking using as a query pattern the leaf expansion period for each species. In the three top graphs, the horizontal axis represents the normalized sum of the errors from all searches made against the others available years. Therefore, the size of the bar indicates how different it is when compared to all others. For instance, in 7.8(a) and for species A, the graph indicates that 2012 is the outlier, followed by 2004 and 2009.

Figures 7.8(b) and 7.8(c) reveal a similar pattern. On this wise, even years of 2012, 2004, and 2009 being the most different for community masks, this is less remarkable

for species B. The year of 2004 is an outstanding: it is always the most particular, as we have observed in Figure 1.1 with a visual CPM comparison against other years. Indeed, when we have investigated why 2004 is so much different considering only species A, we noticed that year was marked by higher temperatures and pluviosity levels.

Figure 7.7: The graphs (a), (b), and (c) were generated selecting the leaf expansion period and using the respective species' ROI: the horizontal axis represents the normalized error sum from all searches against others years.

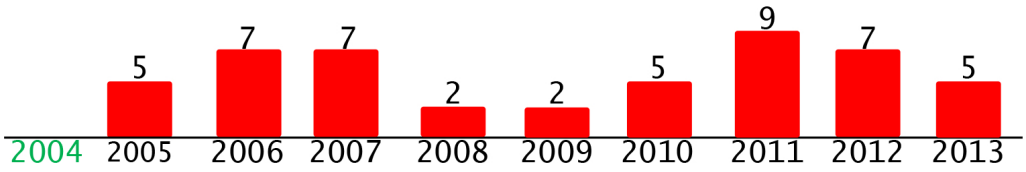


(a) Species A (b) Species B (c) Species C

Source: Author.

Another interesting result revealed is the wave pattern in the leaf expansion period of Figure 7.8. In this case, we used the year 2004 (the first in the series) as the reference. The bar height indicates the number of days when the same pattern appears, before or after the query pattern. For instance, if the query pattern happened on a given day of 2004, the most similar pattern happened five days later on 2005, seven days later on 2006, and so forth. This pattern indicates a characteristic of the plant growing cycle, either appearing later.

Figure 7.8: The graph results from a Leaf-Expansion of 2004 pattern search, using moving window approach, made to the year of 2004. Bar heights represents the number of days that the result happened before or after the query pattern.



Source: Author.

7.3 Scenarios Evaluation

This section evaluates three specific scenarios of the PhenoVis usage.

7.3.1 h_{cc} approach features

Our approach allows the identification of the main colors involved in each events. This technique has restrict palettes that are generated based on the same color scheme of it (HSV). Since these palettes have a better semantic representation, this feature guarantees a quick understanding of the phenomena. For instance, if some color amount is increasing

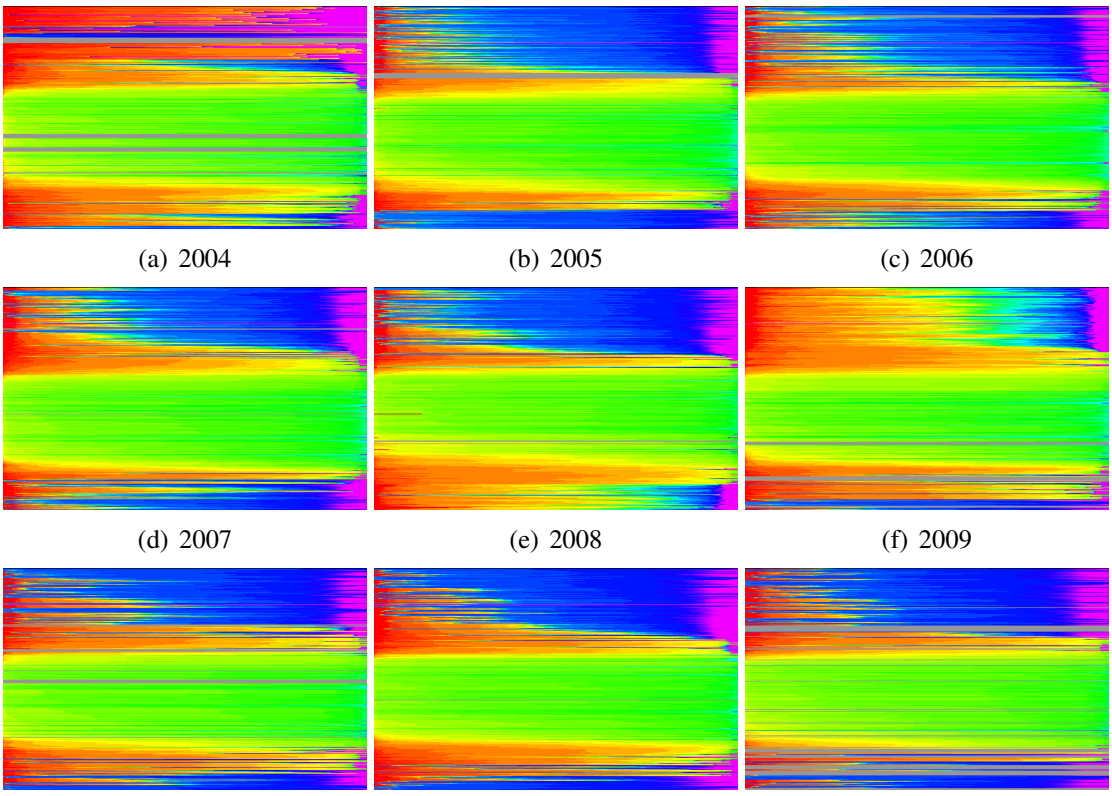
along the time, its color related areas are being more exposed to the camera as well. Some colors meaning in our datasets are:

Green shades represents green pixels increase in the image such as leaf-expansion of the plants.

Blue shades represents snow and rain appearance.

Red shades represents the ground(that, in fact, is often brown).

Figure 7.9: CPMs for the years 2004-2012 using the h_{cc} index. Comparing the results, we observe that the years of 2004 and 2009 are clearly distinct from the others. This result is also confirmed by the general-rank analysis discussed later.



(a) 2004 (b) 2005 (c) 2006

(d) 2007 (e) 2008 (f) 2009

(g) 2010 (h) 2011 (i) 2012

Source: Author.

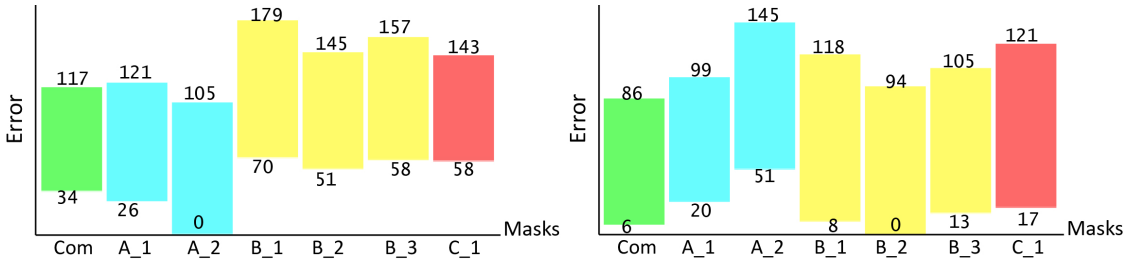
The side-by-side view of CPMs from different years presents relevant observations about the data. In Figure 7.9 we give CPMs for the years 2004-2012. The CPM analysis shows clearly that the years of 2004 and 2009 are distinct from the others. The year of 2004 had the highest temperature and humidity indexes of all years. On the other hand, 2009 had the lowest humidity of all years and lowest snow index, which reveals more of the terrain around trees during winter time. The h_{cc} index was used because its natural color associations allows a more precise identification of such aspects.

7.3.2 Comparing species phenological patterns

One of the primary proposed solution by PhenoVis is the species' comparisons. Comparisons among distinct masks, and species quantity estimations about the community

are supported using General-rank visualization. For instance, if the community mask has small error values to a given species, we can assume that this specie has a significant presence in the community. On the other hand, high similarity errors, represent small presence. In addition, when species are compared with each other, the resulted error estimates how similar two species are. This similarity can be in relation to their events period (if used fixed window), or in relation to the pigment levels involved in the event (if used moving window).

Figure 7.10: Canopies comparison. Each plot summarizes the results of the general-rank search. Each canopy is defined by a mask and analyzed using MSE as the similarity metric and the moving window approach. The range of maximum and minimum errors from the comparison of two canopies masks is shown above and below each bar. Bar colors identify each canopy species. The "Com", in green, abbreviates the community mask. Samples from canopies masks from the same species have minimum error value closer to zero as expected.



(a) Canopy A_2 mask (b) Canopy B_2 mask

Source: Author.

In Figure 7.10, we illustrate how species similarity can be done using ranking results. Using one canopy of species B (canopy B_2), the ranking is generated using the minimum error. Obviously, the other two canopies of B are closer to B_2 than the canopies of A and C. Also, observe how much closer the canopies of B are to the community, which may mean that the community might have a significant number of B canopies.

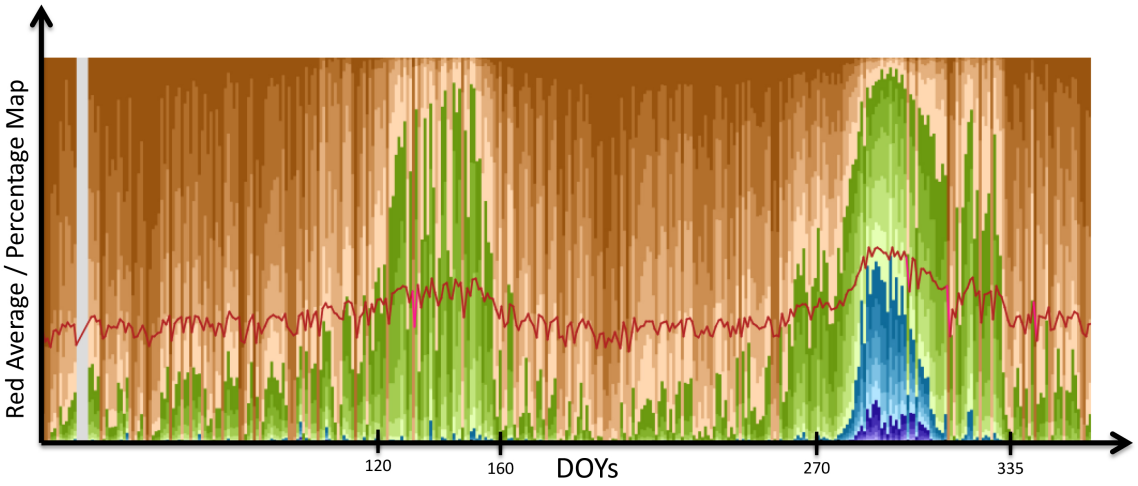
7.3.3 Event detailing

The CPM generated for the year 2006 against the corresponding average plot based on the red chromatic coordinate (displayed as a red line over the CPM) is shown by Figure 7.11. While the average has changed over time, as can be seen between days 120–160 and 270–300, it fails to illustrate their composition as CPM does. The influence of low red intensities (represented by the blue colors) in the change of the second period, for example, can be easily spotted in PhenoVis' CPM.

In the case of using r_{cc} vegetation index, the main highlighted events are leaf expansion and leaf fall. Supposing scenarios where a biologist is analyzing the Figure 7.11. Looking for the r_{cc} line plot, the expert can barely define when the two main events happened. However, two peaks are easily seen by the CPM visualization that was generated based on the same images set than the line plot.

CPMs created using a classificatory palette (such as the shown in Figure 7.11) shows a range of details about each peak, allowing theirs distinction and explanation. Left to right, the first peak refers to leaf expansion. This phenomenon major characteristic is the green pigments increasing in the leafs, which is reflected by indexes such as g_{cc}. However, even on a smaller scale, red pigments also are added during this phase. The second peak refers to the leaf fall. In this event, green pigments levels fall dramatically, whereas the red pigments maintain stable for a while until the leafs fall indeed. This sudden disappearance of the green levels that were occluding the red pigments creates a curl in the CPM that displays higher percentage map channels zones, in this case: purple and blue zones.

Figure 7.11: CPM and the corresponding red average (red line): we can see how the CPM shows much more details. This CPM has been generated using r_{cc} to relative data from the year of 2006.



Source: Author.

8 DISCUSSION AND LIMITATIONS

There are several lessons we learned in the development of this work. While the initial idea of CPMs came early in the development, finding the best form of coloring and display it took many iterations. Since the color interpret is a perceptive task and our visual mining goal variates, it is hard to determine which color combination better support the visual mining. Despite defining our favorites palettes (Palette_1, Palette_2 and HSV cylinder), we implement PhenoVis to accept any palette input.

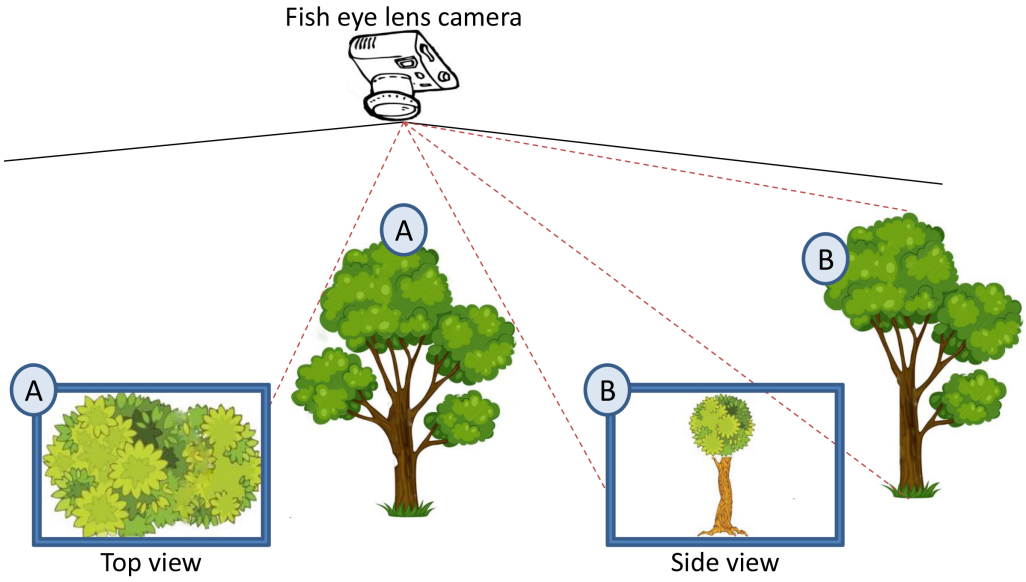
Another difficulty of the CPMs generation was the handling of outliers and missing items. To this problematic items, we have two approaches: ignore or threat them. In the first approach, we can just removed them from consideration, which introduced missing gaps in the information. On the other hand, to threat them we calculate an average using the neighbors values. Next, we generate an artificial value to represent the problematic sample. Analyzing the results we could be sure that these changes presented modestly different results. However, we are not secure of which is semantically the best approach as well as if a third approach could be more significant.

Another limitation is the IOI specification. Besides using pre-determined dates from events (e.g. Leaf Expansion) defined by (NAGAI et al., 2011), the user can also set the starting and ending dates. However, an automatic suggestions of IOI range could improve and speed-up the analysis. This could be done by using, for example, a box-plot (WILLIAMSON; PARKER; KENDRICK, 1989) of the distribution.

The current prototype considers only a single image for each day (at noon), since it allows a fair comparison of different days and avoids lighting variations due to shadows. However, there are other datasets that take multiple images per day, and they can be used for more detailed analysis, such as better species identification. This would be something interesting to incorporate in the prototype.

Figure 8.1 shows different images extraction angles can lead phenological approaches to problematic evaluation. Even covering a wide area range, fish-eye cameras, such as used in TKY, have different distance and angle from each canopy in the area. Canopies samples located distant from the center of the camera have shown smaller similarity than the ones disposed of in the center. In other words, trees disposed far from the center of the camera presents outstanding results. However, this is not an issue only from PhenoVis. Even doing an average plot, the usual approach to this type of data, the samples arranged far from the center have shown a different behavior from others which are not. We consider these problematic samples. We supposed that these estrange behavior happens because these canopies are far from the camera. Elements such as others trees canopies and even trunks on the way can occlude the query canopy during some events (e.g. Leaf-Expansion).

Figure 8.1: This image displays approximately the angles involved in the trees monitoring. As can be observed, (A) images result in a better view from the leafs than the (B) images. This happens because of the distance from the center of the camera. (B) Images are often analyzed with brown shades from the trunk, or even with some other tree in the front of it. Whereas (A), have the biggest range of leafs analyzed, and without no temporal occlusion.



Source: Author.

9 CONCLUSIONS AND FUTURE WORK

Phenology knowledge is important to better understand our environment and its influences in our lives. PhenoVis is a framework for the visual phenological analysis of forest ecosystems that allows us to understand better and interpret features from the forest environment that can lead us to significant conclusions. The carbon balance of the trees and its main components (soil, trees) are displayed in a CPM. Since comparing the resulted CPMs from all years, we enable the understanding of which certain periods may be affected by climate variation. Our approach also provides detailed measurements for the development of the monitored area highlighting the forest growth and productivity. This feature can be used to assess the climate change possible impacts.

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. Chronological percentage maps (CPM), a novel representation presented by this work, 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 CPMs can bring to the analysis and how it can be used to identify species. The evaluation showed how automatic pattern searches can ease the detection of phenological singularities related to weather variations.

As future work, we intend to automatically detect phenological patterns. As of today, such process is completely manual: the user informs start and end dates of the query pattern. Automatic suggestions can improve the analysis using, for example, a box-plot (WILLIAMSON; PARKER; KENDRICK, 1989) of the distribution. Another possibility is to use more images per day, decreasing lighting variations and artifacts. Some datasets already provide such detailed information. We also plan to investigate the integration of the CPM representation with machine learning techniques to perform automatic species identification.

REFERENCES

AHRENDS, H. et al. Tree phenology and carbon dioxide fluxes: use of digital photography for process-based interpretation at the ecosystem scale. **Proceeding of Climate Research**, v.39, n.3, p.261–274, 2009.

AIGNER, W. et al. **Visualization of Time-Oriented Data**. 1st ed. 2011. 286p. (Human-Computer Interaction).

ALBERS, D.; DEWEY, C.; GLEICHER, M. Sequence Surveyor: leveraging overview for scalable genomic alignment visualization. **Proceeding of Visualization and Computer Graphics, IEEE Transactions on**, v.17, n.12, p.2392–2401, Dec 2011.

ALBERTON, B. et al. Using Phenological Cameras to Track the Green Up in a Cerrado Savanna and its On-the-Ground Validation. **Proceeding of Ecological Informatics**, v.19, p.62–70, 2014.

ALMEIDA, J. et al. Remote phenology: applying machine learning to detect phenological patterns in a cerrado savanna. In: E-SCIENCE (E-SCIENCE), 2012 IEEE 8TH INTERNATIONAL CONFERENCE ON, 2012. **...** 2012. p.1–8.

ALMEIDA, J. et al. Plant Species Identification with Phenological Visual Rhythms. In: PROCEEDING OF ESCIENCE (ESCIENCE), 2013 IEEE 9TH INTERNATIONAL CONFERENCE ON, 2013. **...** 2013. p.148–154.

ALMEIDA, J. et al. Visual rhythm-based time series analysis for phenology studies. In: PROCEEDING OF IMAGE PROCESSING (ICIP), 2013 20TH IEEE INTERNATIONAL CONFERENCE ON, 2013. **...** 2013. p.4412–4416.

ALMEIDA, J. et al. Applying machine learning based on multiscale classifiers to detect remote phenology patterns in cerrado savanna trees. **Ecological Informatics**, v.23, p.49–61, 2014.

ANDRIENKO, G.; ANDRIENKO, N. Spatio-temporal aggregation for visual analysis of movements. In: IEEE SYMPOSIUM ON VAST, 2008. **...** 2008. p.51–58.

BOYANDIN, I. et al. Flowstrates: an approach for visual exploration of temporal origin-destination data. In: PROCEEDING OF PROCEEDINGS OF THE 13TH EUROGRAPHICS / IEEE - VGTC CONFERENCE ON VISUALIZATION, 2011, Aire-la-Ville, Switzerland, Switzerland. **...** Eurographics Association, 2011. p.971–980. (EuroVis'11).

BRADLEY, E.; ROBERTS, D.; STILL, C. Design of an image analysis website for phenological and meteorological monitoring. **Proceeding of Environ. Modelling and Soft.**, v.25, n.1, p.107–116, 2010.

CRIMMINS, M. A.; CRIMMINS, T. M. Monitoring plant phenology using digital repeat photography. **Proceeding of Environmental management**, v.41, n.6, p.949–958, 2008.

CUI, W. et al. Let It Flow: a static method for exploring dynamic graphs. In: IEEE PACIFIC VISUALIZATION SYMPOSIUM (PACIFICVIS), 2014. **...** 2014. p.121–128.

EERENS, H. et al. Image time series processing for agriculture monitoring. **Environmental Modelling & Software**, v.53, p.154–162, 2014.

FUNG, K. **Numbers Rule Your World**: the hidden influence of probabilities and statistics on everything you do. 2010.

GILLESPIE, A. R.; KAHLE, A. B.; WALKER, R. E. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. **Proceeding of Remote Sensing of Environment**, v.22, n.3, p.343 – 365, 1987.

GRANADOS, J. A. et al. EcoIP: an open source image analysis toolkit to identify different stages of plant phenology for multiple species with pan-tilt-zoom cameras. **Proceeding of Ecological Informatics**, v.15, p.58–65, 2013.

GRATZL, S. et al. LineUp: visual analysis of multi-attribute rankings. **Proceeding of IEEE Trans. on Vis. and Comp. Graphics**, v.19, n.12, p.2277–2286, Dec 2013.

HADLAK, S. et al. Supporting the Visual Analysis of Dynamic Networks by Clustering associated Temporal Attributes. **Proceeding of IEEE Trans. on Vis. and Comp. Graphics**, v.19, n.12, p.2267–2276, Dec 2013.

IDE, R.; OGUMA, H. Use of digital cameras for phenological observations. **Ecological Informatics**, v.5, n.5, p.339–347, 2010.

JIE LI KANG ZHANG, Z.-P. M. Vismate: interactive visual analysis of station-based observation data on climate changes. In: IEEE SYMPOSIUM ON VAST, 2014. **...** 2014.

KEIM, D. A. Information Visualization and Visual Data Mining. **IEEE Trans. on Vis. and Comp. Graph.**, Piscataway, NJ, USA, v.8, n.1, Jan. 2002.

KEIM, D. A. et al. Pixel bar charts: a visualization technique for very large multi-attribute data sets. **Proceeding of Information Visualization**, v.1, n.1, p.20–34, Mar. 2002.

LEE, D. W. et al. Pigment dynamics and autumn leaf senescence in a New England deciduous forest, eastern USA. **Proceeding of Ecological Research**, v.18, n.6, 2003.

LIN, S. et al. Selecting Semantically-Resonant Colors for Data Visualization. **Proceeding of Computer Graphics Forum**, v.32, n.3pt4, p.401–410, 2013.

M., W. D. et al. Color indices for weed identification under various soil, residue, and lighting conditions. **Proceeding of Transactions of the ASAE**, v.38, n.1, p.259–269, 1995.

MENZEL, A. et al. European phenological response to climate change matches the warming pattern. **Proceeding of Global Change Biology**, v.12, n.10, p.1969–1976, 2006.

MORISETTE, J. et al. Tracking the rhythm of the seasons in the face of global change: phenological research in the 21st century. **Proceeding of Frontiers in Ecology and the Environment**, v.7, n.5, p.253–260, June 2009.

NAGAI, S. et al. Using digital camera images to detect canopy condition of deciduous broad-leaved trees. **Proceeding of Plant Ecology & Diversity**, v.4, n.1, 2011.

NEGI, G. Leaf and bud demography and shoot growth in evergreen and deciduous trees of central Himalaya, India. **Proceeding of Trees**, v.20, n.4, p.416–429, 2006.

NGO, C.; PONG, T.; CHIN, R. Detection of gradual transitions through temporal slice analysis. In: COMPUTER VISION AND PATTERN RECOGNITION, 1999. IEEE COMPUTER SOCIETY CONFERENCE ON., 1999. **...** 1999. v.1, p.–41 Vol. 1.

NGUYEN, K. T.; ROPINSKI, T. Large-scale multiple sequence alignment visualization through gradient vector flow analysis. In: PROCEEDING OF BIOLOGICAL DATA VISUALIZATION (BIOVIS), 2013 IEEE SYMPOSIUM ON, 2013. **...** 2013.

NISHIDA, K. Proceeding of. Phenological Eyes Network (PEN)–a validation network for remote sensing of the terrestrial ecosystems. **AsiaFlux Newsletter**, v.21, p.9–13, 2007.

OLIVEIRA, G. et al. Visualizing Running Races through the Multivariate Time-Series of Multiple Runners. , p.99–106, Aug 2013.

SCHWARTZ, M. D. et al. 5. **Phenological Data, Networks, and Research**: north america. 2nd ed. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2013.

POLGAR, C. A.; PRIMACK, R. B. Leaf-out phenology of temperate woody plants: from trees to ecosystems. **Proceeding of New Phytologist**, v.191, n.4, p.926–941, 2011.

REICH, P. B. Phenology of Tropical Forests: patterns, causes and consequences. **Proceeding of Canadian Journal of Botany**, v.73, p.164–174, 1995.

RICHARDSON, A. D. et al. Use of Digital Webcam Images to Track Spring Greep-up in a Deciduous Broadleaf Forest. **Proceeding of Oecologia**, v.152, p.323–334, 2007.

RICHARDSON, A. D. et al. Near-surface remote sensing of spatial and temporal variation in canopy phenology. **Proceeding of Ecological Applic.**, v.19, n.6, p.1417–1428, 2009.

ROTZER, T.; GROTE, R.; PRETZSCH, H. The Timing of Bud Burst and Its Effect on Tree Growth. **Proceeding of International Journal of Biometeorology**, v.48, p.109–118, 2004.

S, T. et al. Phenological Eyes Network for validation of remote sensing data. **Proceeding of Remote Sensing Society of Japan**, v.25, p.282 288, 2005.

SCHNORR, L.; LEGRAND, A. Visualizing More Performance Data Than What Fits on Your Screen. In: CHEPTSOV, A. et al. (Ed.). **Tools for High Performance Computing 2012**. 2013. p.149–162.

SONNENTAG, O. et al. Digital repeat photography for phenological research in forest ecosystems. **Agricultural and Forest Meteorology**, v.152, p.159–177, 2012.

UDELHOVEN, T. TimeStats: a software tool for the retrieval of temporal patterns from global satellite archives. **Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of**, v.4, n.2, p.310–317, June 2011.

WALTHER, G.-R. et al. Ecological responses to recent climate change. **Proceeding of Nature**, v.416, n.6879, p.389–395, Mar. 2002.

WANG, Z. et al. Visual Traffic Jam Analysis Based on Trajectory Data. **Proceeding of IEEE Trans. on Vis. and Comp. Graph.**, v.19, n.12, p.2159–2168, 2013.

WILLIAMSON, D. F.; PARKER, R. A.; KENDRICK, J. S. The box plot: a simple visual method to interpret data. **Proceeding of Annals of internal medicine**, v.110, n.11, p.916–921, 1989.

WOOD, J. et al. BallotMaps: detecting name bias in alphabetically ordered ballot papers. **Proceeding of IEEE Trans. on Vis. and Comp. Graphics**, v.17, n.12, p.2384–2391, Dec 2011.

ZHANG, X. et al. Long-term detection of global vegetation phenology from satellite instruments. **Proceeding of Phenology and Climate Change**, v.16, p.297–320, 2012.