José Carlos Conti

Eficácia de medidas de similaridade para a classificação de séries temporais associadas ao comportamento fenológico de plantas

Limeira, 2013

UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE TECNOLOGIA

José Carlos Conti

Eficácia de medidas de similaridade para a classificação de séries temporais associadas ao comportamento fenológico de plantas

Dissertação apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Tecnologia.

Orientador: Prof. Dr. Luiz Camolesi Júnior
Co-orientador: Prof. Dr. Ricardo da Silva Torres

Exemplar corresponde à redação final da dissertação, devidamente corrigida, defendida por José Carlos Conti e aprovada pela Banca Examinadora.

Orientador:

Data:

Limeira, 2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Tecnologia
Vanessa Evelyn Costa - CRB 8/8295

C767e

Conti, José Carlos, 1966-
Eficácia de medidas de similaridade para a classificação de séries temporais associadas ao comportamento fenológico de plantas. / José Carlos Conti. – Limeira, SP : [s.n.], 2013.

Orientador: Luiz Camolesi Júnior.
Coorientador: Ricardo da Silva Torres.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Tecnologia.

1. Mineração de dados (Computação). 2. Similaridade (Geometria). 3. Análise de séries temporais. 4. Plantas - classificação. 5. Fenologia. I. Camolesi Júnior, Luiz. II. Torres, Ricardo da Silva. III. Universidade Estadual de Campinas. Faculdade de Tecnologia. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Eficácia de medidas de similaridade para a classificação de séries temporais associadas ao comportamento fenológico de plantas.
Palavras-chave em inglês:
Data mining (Computing)
Similarity (Geometry)
Time series analysis
Plants - classification
Phenology
Área de concentração: Tecnologia e Inovação
Titulação: Mestre em Tecnologia
Banca examinadora:
Luiz Camolesi Júnior [Orientador]
Ana Estela Antunes da Silva
Leonor Patrícia Cerdeira Morellato
Data de defesa: 11-12-2013
Programa de Pós-Graduação: Tecnologia

DISSERTAÇÃO DE MESTRADO EM TECNOLOGIA

ÁREA DE CONCENTRAÇÃO: TECNOLOGIA E INOVAÇÃO

Eficácia de medidas de similaridade para a classificação de séries temporais associadas ao comportamento fenológico de plantas

José Carlos Conti

A Banca Examinadora composta pelos membros abaixo aprovou esta Dissertação:

Prof. Dr. Luiz Camolesi Junior
UNICAMP/FT
Presidente

Profa. Dra. Ana Estela Antunes da Silva
UNICAMP/FT

Profa. Dra. Leonor Patricia Cerdeira Morellato
UNESP

Resumo

Fenologia é o estudo de fenômenos naturais periódicos e sua relação com o clima. Nos últimos anos, tem se apresentado relevante como o indicador mais simples e confiável dos efeitos das mudanças climáticas em plantas e animais. É nesse contexto que se destaca o *e-phenology*, um projeto multidisciplinar envolvendo pesquisas na área de computação e fenologia. Suas principais características são: o uso de novas tecnologias de monitoramento ambiental, o fornecimento de modelos, métodos e algoritmos para apoiar o gerenciamento, a integração e a análise remota de dados de fenologia, além da criação de um protocolo para um programa de monitoramento de fenologia. Do ponto de vista da computação, as pesquisas científicas buscam modelos, ferramentas e técnicas baseadas em processamento de imagem, extraindo e indexando características de imagens associadas a diferentes tipos de vegetação, além de se concentrar no gerenciamento e mineração de dados e no processamento de séries temporais. Diante desse cenário, esse trabalho especificamente, tem como objetivo investigar a eficácia de medidas de similaridade para a classificação de séries temporais sobre fenômenos fenológicos caracterizados por vetores de características extraídos de imagens de vegetação. Os cálculos foram realizados considerando regiões de imagens de vegetação e foram considerados diferentes critérios de avaliação: espécies de planta, hora do dia e canais de cor. Os resultados obtidos oferecem algumas possibilidades de análise, porém na visão geral, a medida de distância *Edit Distance with Real Penalty* (ERP) apresentou o índice de acerto mais alto com 29,90%. Adicionalmente, resultados obtidos mostram que as primeiras horas do dia e no final da tarde, provavelmente devido à luminosidade, apresentam os índices de acerto mais altos para todas as visões de análise.

Palavras chave: Mineração de dados (Computação), Similaridade (Geometria), Análise de séries temporais, Plantas - classificação, Fenologia.

Abstract

Phenology is the study of periodic natural phenomena and their relationship to climate. In recent years, it has gained importance as the more simple and reliable indicator of effects of climate changes on plants and animals. In this context, we emphasizes the e-phenology, a multidisciplinary research project in computer science and phenology. Its main characteristics are: The use of new technologies for environmental monitoring, providing models, methods and algorithms to support management, integration and remote analysis of data on phenology, and the creation a protocol for a program to monitoring phenology. From the computer science point of view, the e-phenology project has been dedicated to creating models, tools and techniques based on image processing algorithms, extracting and indexing image features associated with different types of vegetation, and implementing data mining algorithms for processing time series. This project has as main goal to investigate the effectiveness of similarity measures for the classification of time series associated with phenological phenomena characterized by feature vectors extracted from images. Conducted experiments considered different regions containing individuals of different species and considering different criteria such as: plant species, time of day and color channels. Obtained results show that the *Edit Distance with Real Penalty* (ERP) distance measure yields the highest accuracy. Additionally, the analyzes show that in the early morning and late afternoon, probably due to light conditions, it can be observed the highest accuracy rates for all views analysis.

Keywords: Data mining (Computing), Similarity (Geometry), Time series analysis, Plants - classification, Phenology.

Sumário

Agradecimento

Agradeço primeiramente a Deus, que está presente em minha vida.

Agradeço à minha querida esposa Alessandra que sempre me apoia e entende a importância deste projeto.

Agradeço à minha querida filha Isabella que também me apoia e é a grande inspiração para a minha vida.

Agradeço à minha mãe pela paciência na minha ausência em alguns momentos de datas comemorativas em família.

Agradeço ao meu pai, que mesmo não mais entre nós, está olhando por mim.

Agradeço aos meus irmãos e irmãs, cunhados e cunhadas, sobrinhos e sobrinhas que entenderam o meu esforço e acreditaram em mim.

Agradeço ao meu sogro e sogra pela compreensão e paciência.

Agradeço ao Professor Doutor Luiz Camolesi Jr., por toda orientação e apoio, muitas vezes fundamental para o meu aprendizado.

Agradeço ao Professor Doutor Ricardo da Silva Torres, que sempre me apoiou e permitiu o meu ingresso como aluno especial na primeira disciplina cursada no Instituto de Computação.

Agradeço também aos demais docentes do Instituto de Computação e da Faculdade de Tecnologia, a Professora Doutora Ana Estela Antunes da Silva e a Professora Doutora Leonor Patrícia Cerdeira Morellato e os membros suplentes Professor Doutor Guilherme Palermo Coelho e Professor Doutor João Eduardo Ferreira da banca examinadora.

Agradeço ainda aos amigos que direta ou indiretamente me ajudaram e apoiaram.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

DTW - *Dynamic Time Warping*
EDR - *Edit Distance on Real sequence*
ERP - *Edit distance with Real Penalty*
EVI - *Enhanced Vegetation Index*
FAPESP - Fundação de Amparo à Pesquisa do Estão de São Paulo
KDD - *Knowledge Discovery in Databases*
KNN - *K-Nearest Neighbor*
L1 - Distância de Manhatan
L2 - Distância Euclidiana
LCSS - *Longest Common Subsequence*
MODIS - *Moderate Resolution Imaging Spectroradiometer*
NDVI - *Normalized Difference Vegetation Index*
RECOD - *Recognition Database*
ROI - *Region of Interest*
SGBDs - Sistemas Gerenciadores de Bancos de Dados
TIDES - *Time Series Oscillation Descriptor*
UNESP - Universidade Estadual Paulista
UNICAMP - Universidade Estadual de Campinas
ZNCC - *Zero-mean Normalized Cross Correlation*

1 Introdução

O uso de radares meteorológicos, ou sensores eventualmente remotos, têm aumentado significativamente nos últimos anos, gerando grandes volumes de dados tais como, temperatura e precipitação. Além disso, os modelos de mudanças climáticas têm sido processados para diversos cenários, contribuindo para o aumento da quantidade de dados climáticos (MCCLOY, 2010; HUDSON et al., 2009). A análise desses dados torna-se um desafio muito importante para os pesquisadores (UOTILA et al., 2007; SCHNEIDER et al., 2009), uma vez que as mudanças climáticas podem influenciar no ciclo de vida de plantas e animais.

Uma representação dessa análise surge no contexto dos estudos de fenologia. Define-se fenologia como o estudo de fenômenos naturais recorrentes e sua relação com o clima (SCHWARTZ, 2003). É uma ciência dedicada a observar os ciclos (fenofases) de plantas e animais e sua relação com os dados meteorológicos locais, bem como para interações bióticas e filogenia (RATHCKE & LACEY, 1985). Dentre as principais fenofases de plantas destacam-se brotamento, floração, frutificação e queda foliar.

Do ponto de vista da computação, ainda há poucas pesquisas dedicadas à criação de modelos, ferramentas e técnicas baseadas em processamento de imagem, no sentido de extrair e indexar características extraídas de imagens associadas a mudanças na vegetação. Além disso, identifica-se a necessidade de novas abordagens para o gerenciamento e mineração de dados, assim como para o processamento de séries temporais (ALMEIDA et al., 2012). O estudo de fenologia emprega técnicas de observação remota ou em campo e a geração de séries temporais [1].

Série temporal é definida como um conjunto de observações registradas ao longo do tempo e analisadas sequencialmente (BROCKELL & DAVIS, 1997). É importante compreender que, em séries temporais, as observações com proximidade temporal são dependentes sendo necessário analisar e modelar essa dependência (OLIVEIRA, 2007). Nesse contexto, a análise de séries temporais tem se tornado importante para o processo de tomada de decisão, inclusive na fenologia.

Em relação aos fenômenos fenológicos baseado em séries temporais, há uma demanda por ferramentas e técnicas de mineração que consigam encontrar padrões nos dados coletados (GUYON et al., 2010). Nesse sentido, o desafio é extrair padrões relevantes sobre os conjuntos de dados, analisando as variáveis de maneira integrada e

[1] http://www.recod.ic.unicamp.br/ephenology

indicando interdependências em séries de dados que cobrem longos períodos de tempo (PINTO et al., 2000).

1.1 Objetivo

Considerando a importância da tarefa de classificação nas pesquisas em fenologia, esse trabalho tem como principal objetivo investigar a eficácia de medidas de distância, avaliando a sua similaridade para a classificação de séries temporais sobre fenômenos fenológicos modelados em vetores de características extraídos de imagens de vegetação. A análise de eficácia das medidas de distância consiste em avaliar os percentuais médios de acerto e desvio padrão categorizados por espécie de vegetação, canal de cor (*Red, Green* e *Blue*) e hora do dia.

Nesse trabalho, foram empregados os seguintes algoritmos para calcular as distâncias entre as séries temporais: L1 - Distância de *Manhatan (*YI & FALOUTSOS, 2000*)*; L2 - Distância Euclidiana (FALOUTSOS et al., 1994); DTW - *Dynamic Time Warping* (KEOGH, 2002); LCSS - *Longest Common Subsequence (*VLACHOS et al., 2002*)*; ZNCC - *Zero-mean Normalized Cross Correlation (*MARTIN & CROWLEY, 1995*)*; EDR - *Edit Distance on Real Sequence* (CHEN et al., 2005) e; ERP - *Edit Distance with Real Penalty (*CHEN et al., 2004*).* O classificador KNN (K-*Nearest Neighbor*) foi usado para ordenar estes cálculos e obter as menores distâncias destas séries.

Para atender o objetivo foram realizadas análises de identificação de indivíduos da mesma espécie de vegetação pertencentes às classes de teste em relação à classe de treino, confrontando-as considerando características semelhantes do ponto de vista de segmentação de cor (*Red, Green* e *Blue*), e se há variação ao longo do dia.

1.2 Projeto Relacionado

Este trabalho está relacionado ao projeto *e-phenology* que envolve pesquisas na área de computação e fenologia. Este projeto foi iniciado em 2011 sendo desenvolvido por cientistas da computação do Laboratório RECOD (UNICAMP) e pesquisadores em fenologia do Laboratório de Fenologia (UNESP) de Rio Claro. O projeto é financiado pelo Instituto Virtual FAPESP-Microsoft e tem como principais características o uso de novas tecnologias de monitoramento ambiental, o fornecimento de modelos, métodos e algoritmos para suportar gerenciamento, integração e análise remota de dados de

fenologia, além da criação de um protocolo para um programa de monitoramento de fenologia [2] [3].

As principais contribuições do projeto *e-phenology* para a área de fenologia são: criar modelos e metodologia de análise de mudanças climáticas baseada na exploração de novos índices de fenologia; para a área de computação, criar modelos, ferramentas e técnicas baseadas em processamento de imagem, visando extrair e indexar características de imagens associadas a tipos de vegetação; e em bancos de dados, gerenciar e minerar dados, processar e anotar séries temporais (MORELLATO & TORRES, 2011).

1.3 Motivação e Justificativa

Existem pesquisas que abordam técnicas de mineração de dados usando padrões e séries temporais, como o trabalho de ROMANI et al., (2010). Poucos trabalhos propostos tratam do problema de se comparar a eficácia de medidas de similaridade (DING et al., 2008; ESLING & AGON, 2012) em tarefas de mineração de séries temporais, bem como não buscam correlacionar os dados de uma série com outras séries temporais no sentido de enriquecer a análise desses dados, ou seja, não se concentram em identificar se os dados de uma série podem influenciar nos dados de outras séries (MUEEN et al., 2010).

Considerando a complexidade do projeto *e-phenology* e o volume de informações que se pretende gerenciar, armazenar e analisar, este trabalho contribui com a investigação de medidas de distância, avaliando a eficácia dessas medidas para a classificação de séries temporais. Exemplos de alguns desafios para a área de fenologia incluem a similaridade entre séries temporais fenológicas de espécies diferentes.

1.4 Materiais e Métodos

Os dados usados para a realização das avaliações foram obtidos do sistema de monitoramento remoto para captura de informações de fenologia localizado em uma torre a 18 Metros de altura no Cerrado Stricto Sensu, vegetação na savana localizada em Itirapina (22º10'49.18"S / 47º52'16.54"O), no Estado de São Paulo, Brasil (ALMEIDA

[2] http://www.recod.ic.unicamp.br/ephenology (Último acesso em outubro de 2013).
[3] http://www.fapesp.br/publicacoes/microsoft/microsoft_morellato.pdf

et al., 2012). Uma câmera digital controlada por um sensor e alimentada por bateria solar de 12 volts está posicionada no topo da torre e realiza fotos da área (Figura 1).

Figura 1: Torre de 18 M onde está instalada a câmera digital para registrar fenologia
ALMEIDA et al., 2011

A fonte de dados primária para a realização do trabalho são arquivos com características de imagens que foram registradas no período de 29 de agosto a 03 de outubro de 2011, considerando imagens JPEG (1280 x 960 pixels de resolução) por hora, das 06:00 às 18:00 horas.

Estas imagens foram fornecidas pelo departamento de Biologia da Unesp de Rio Claro em parceria com o Laboratório RECOD da Unicamp e consiste em imagens tiradas na região do cerrado. A Figura 2 mostra um exemplo de imagem obtida.

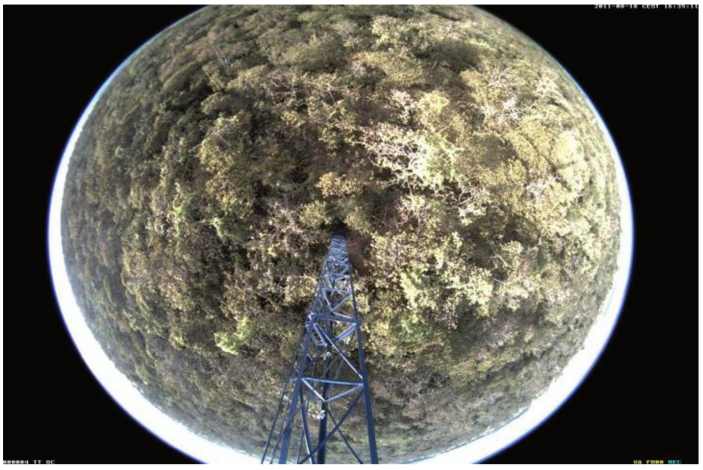

Figura 2: Imagem registrada pela Câmera Digital

A análise de uma imagem define diferentes regiões de interesse, conforme proposto por RICHARDSON et al., (2007). A partir de uma imagem (foto da vegetação) e uma máscara (imagem binária em preto e branco) que demarca uma região nessa imagem (pode ser toda a vegetação ou uma espécie específica), os valores *meanR*, *meanG* e *meanB* representam para cada canal de cor primária, *Red*, *Green* e *Blue* respectivamente, a média dos valores de intensidade de todos os pixels da imagem demarcados pela máscara, ou seja, a cor média da região. Os valores relR, relG e relB representam a fração relativa de contribuição de cada canal de cor e é dado por:

$$rel\ \{R,G,B\} = mean\{R,G\ ou\ B\} \div (mean\ R + mean\ G + mean\ B)$$

Usando os arquivos de imagens, ocorre um processo para obtenção dos arquivos de vetores de características que são usados neste trabalho para a avaliação das medidas de distância. O processo para a obtenção destes arquivos de vetores de características é usado no projeto *e-phenology* para a realização das pesquisas. Este trabalho, da mesma forma, aproveitou-se deste processo utilizando-se diretamente vetores de características produzidos.

A Figura 3 apresenta resumidamente as fases do processo, sendo:

A) Imagens são capturadas pela câmera digital que fica localizada na torre;

B) Os valores de RGB são calculados de modo que os tornem segmentados;

C) Algoritmos são empregados para a extração de séries temporais de regiões da imagem, onde são representados o percentual de cor, os dias em que foram monitorados considerando os horários entre 6 da manhã e 6 da tarde.

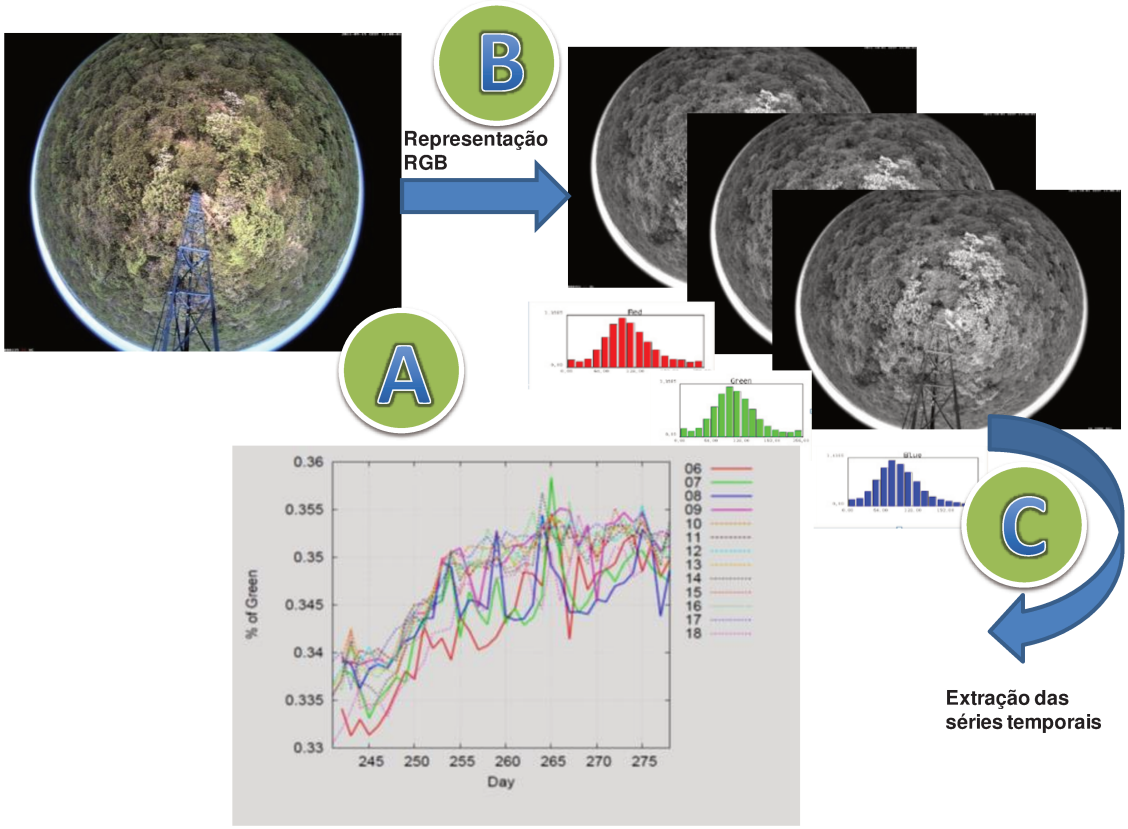

Figura 3: Representação do estudo de fenologia - Extraída de: http://www.recod.ic.unicamp.br/ephenology/

Os volumes de dados primários para a realização deste trabalho estão estruturados da seguinte forma:

1. Para cada hora e cada canal de cor, existem 8813 séries temporais.
2. São 13 horas (6:00 as 18:00 horas) e 3 canais de cor (*Red, Green*, e *Blue*).
3. Desse modo, há 8.813 (séries) x 13 (horas) x 03 (canais de cor) = 343.707 registros.
4. O conjunto de teste é formado por registros das espécies *Aspidosperma tomentosum* (63 registros), *Miconia rubiginosa* (30 registros) e *Pouteria ramiflora* (55 registros). O conjunto de treino é formado por registros das espécies *Aspidosperma tomentosum* (35 registros), *Caryocar brasiliensis* (113 registros), *Myrcia guianesis* (27 registros), *Miconia rubiginosa* (24 registros), *Pouteria torta* (8 registros) e *Pouteria ramiflora* (43 registros). Os demais registros (8415) são atribuídos a uma classe desconhecida. Esta classe desconhecida foi desconsiderada na análise devido à diversidade de vegetação na região do cerrado. Essa distribuição ocorre para cada hora e canal de cor.

   O processo de definição das regiões tem as seguintes etapas: Primeiramente, máscaras associadas a indivíduos de espécies de interesse são definidas (usualmente por especialistas em fenologia). Em seguida, as regiões definidas por estas máscaras

são particionadas por um algoritmo de segmentação em sub-regiões menores. Mais detalhes acerca do processo de segmentação podem ser obtidos no trabalho de ALMEIDA et al., (2013).

1.5 Metodologia do Trabalho

As medidas de distância foram usadas para calcular quão similares são as séries temporais associadas às características das imagens de vegetação. Foram consideradas as medidas L1, L2, DTW, LCSS, ZNCC, EDR e ERP. O protocolo considera o uso das medidas de distância em tarefas de classificação envolvendo o classificador K-*Nearest Neighbor* (KNN), largamente utilizado pela comunidade de aprendizado de máquina (KASHYAP et al., 2011). O objetivo é classificar as regiões de espécies de interesse de acordo com sua proximidade, de acordo com as suas características de cor.

Para a realização da analise de eficácia das medidas de distância, foram empregados os seguintes procedimentos :

1. Primeiramente, calculam-se as distâncias dos registros pertencentes ao conjunto de teste em relação aos registros pertencentes ao conjunto de treino;

2. Na sequência aplica-se o classificador K-NN para determinar a classe de instâncias do conjunto de teste;

3. Por último, avaliam-se as medidas de distância de acordo com resultados de acurácia observados a partir do seu uso no classificador K-NN.

1.6 Organização da Dissertação

Os demais capítulos dessa dissertação estão estruturados da seguinte forma:

- O Capítulo 2 apresenta uma revisão da literatura, analisando alguns conceitos sobre mineração de dados, séries temporais e medidas de similaridade.
- O Capítulo 3 apresenta conceitos e projetos de fenologia.
- O Capítulo 4 apresenta o desenvolvimento do trabalho, as visões de análise e os resultados do ponto de vista de eficácia.
- O Capítulo 5 apresenta a conclusão e possíveis trabalhos futuros.

2 Mineração de Séries Temporais

Este capítulo apresenta uma revisão bibliográfica, descrevendo os conceitos relacionados à mineração de séries temporais.

2.1 Considerações Iniciais

Mineração de dados é um processo responsável por extrair informações de um conjunto de dados disponíveis em uma Base de Dados usando mecanismos diferentes daqueles oferecidos pelos Sistemas Gerenciadores de Bancos de Dados (SGBDs) para consultas (TAN et al., 2005).

Nesse capítulo, será apresentado o processo de descoberta do conhecimento, a definição de séries temporais e algumas medidas de distância utilizadas na literatura.

2.2 Processo de Descoberta do Conhecimento

Mineração de dados é uma área de pesquisa multidisciplinar que envolve tecnologia de Bancos de Dados, inteligência artificial, aprendizagem de máquina, redes neurais, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados (FAYAD et al., 1996).

Para outros pesquisadores, mineração de dados é definida como o processo de extrair conhecimento de grandes volumes de dados (HAN et al., 2001). Esse processo é parte de um contexto maior conhecido como descoberta de conhecimento em Bancos de Dados (*KDD – Knowledge Discovery in Databases*). O *KDD* é composto das seguintes etapas:

1. Entendimento do domínio da aplicação: É o conhecimento relevante na identificação da meta do processo de descoberta do conhecimento do ponto de vista de interesse do usuário.
2. Criação de uma Base de Dados: Define a Base de Dados ou especificação de alguns atributos, na qual a descoberta do conhecimento será aplicada.
3. Limpeza e processamento: Consiste na remoção de ruídos se necessário, selecionando as informações importantes para modelagem, além de utilizar estratégias para aplicar regras em atributos cujos valores são indefinidos ou inexistentes.
4. Representação do dado: Encontrar características úteis para representar o dado dependendo da meta ou tarefa. Com o uso de técnicas de redução da

dimensionalidade, o número de atributos pode ser reduzido, potencialmente impactando a eficiência e eficácia de técnicas de mineração.

5. Aplicação da tarefa de mineração de dados: Baseado na meta do usuário estabelecida na Etapa 1, aplica-se uma das tarefas de mineração de dados como por exemplo em sumarização, classificação, regressão ou agrupamento.

6. Definição do algoritmo: Escolha de um algoritmo de mineração a ser usado para encontrar padrões nos dados.

7. Busca por padrões de interesse: Definição de uma forma de representação dos resultados como regras ou árvores de decisão, regressão ou agrupamento.

8. Interpretação dos padrões de mineração: Possibilidade de retornar para cada etapa anterior para melhor iteração.

9. Ação sobre o conhecimento descoberto: Com base na análise dos resultados, comparar com outros trabalhos e eventualmente analisar se não existem conflitos com os objetivos iniciais propostos.

Esse processo de descoberta do conhecimento é ilustrado na Figura 4.

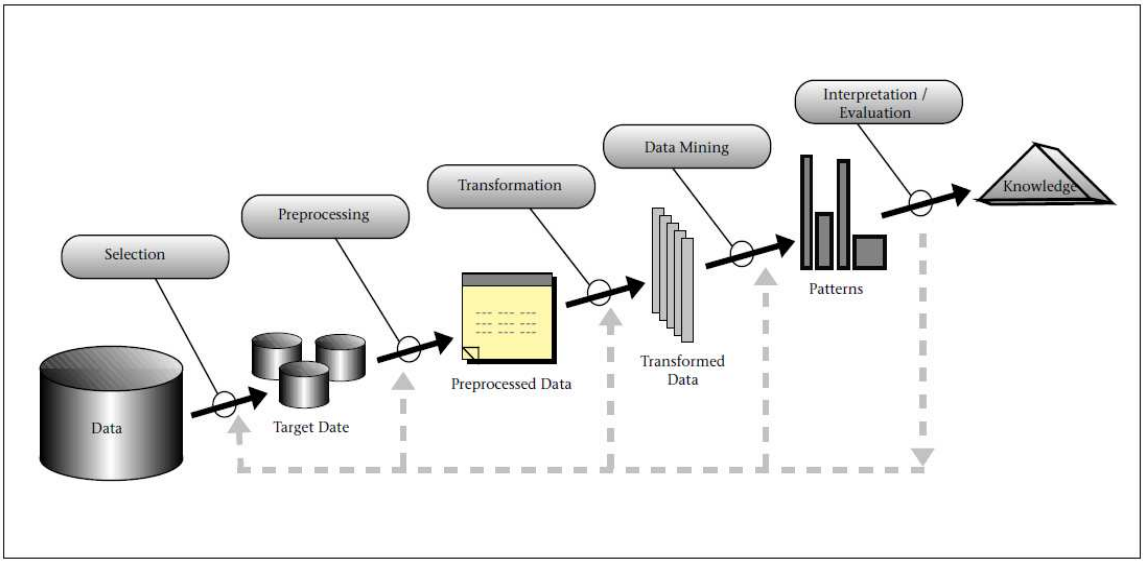

Figura 4: Etapas do KDD - Extraído de (FAYAD et al., 1996)

É importante destacar que nesse processo apresentado podem ocorrer interações entre a etapa corrente e a etapa imediatamente anterior, possibilitando o refinamento da etapa, e consequentemente favorecendo a descoberta do conhecimento em Bases de Dados. Nesse processo, a Mineração de Dados possui uma metodologia própria para preparação e exploração dos dados, interpretação dos resultados e assimilação dos conhecimentos minerados.

2.3 Séries Temporais

No contexto do processo de mineração de dados, a análise de séries temporais se tornou uma importante fonte de investigação em várias áreas (ANDRIENKO et al., 2010; MEZER et al., 2009; OTRANTO, 2010).

Uma série temporal é uma sequência de observações ordenadas no tempo ou no espaço. Os valores de medições se alteram ao longo do tempo, essa variação é representada por uma série temporal (BROCKELL & DAVIS, 1997).

Matematicamente uma série temporal X é representada da seguinte forma: $X = \{X_{t-1}, X_{t-2}, X_{t-3}...\}$, ou seja, a Série Temporal *X* corresponde a um conjunto das medições em relação ao tempo *t.* Existem dois tipos de Séries Temporais: contínuas, em que as observações estão em todos os instantes de tempo (ex: Eletrocardiograma, bolsas de valores); e discretas, em que existe uma observação em espaço de tempo normalmente regulares (ex: Logs de Servidor) , (OLIVEIRA, 2007).

A análise de Séries Temporais consiste na aplicação de modelos matemáticos e estatísticos com o objetivo de quantificar e compreender o fenômeno da variação temporal. Essa análise é efetuada considerando o objetivo de analisar o passado, no sentido de reconhecer conhecimento importante desses dados e, por consequência possibilitar a predição do futuro, tentando através da análise dos dados, construir um modelo que permite prever a evolução futura da série temporal (OLIVEIRA, 2007).

Uma série temporal representa uma coleção de valores obtidos de medições sequenciais ao longo do tempo (ESLING & AGON, 2012). Mineração de dados de séries temporais deriva no desejo de materializar nossa habilidade natural para visualizar/identificar padrões de interesse. Os pesquisadores dependem do uso de esquemas complexos, tais como: tarefas, técnicas e algoritmos de mineração, no sentido de buscar realizar suas pesquisas.

2.3.1 Domínio de Dados

A realização do método de busca deve considerar a natureza dos dados, com o propósito de explorar certas propriedades do espaço em que estão inseridos. A maioria dos métodos de busca sugere que os dados pertencem a um espaço n-dimensional, o que permite o uso de propriedades geométricas na solução. Por outro lado, alguns métodos buscam ser menos rigorosos, como o caso dos métodos baseados em espaços métricos, que só usam as distâncias entre os pontos, e nenhuma outra propriedade geométrica (UENO & LEE, 2006).

2.3.1.1. Espaço Métrico

MONARD, (2004) define um espaço métrico como =<*S,d*>=, onde *S* é um universo de elementos e *d* é a função de distância, definida sobre os elementos em *S,* que mede a dissimilaridade entre os objetos e satisfaz as seguintes condições, $\forall x, y, z \in S$:

$$(1)\ d(x,y) \geq 0 \qquad \textit{positividade}$$
$$(2)\ d(x,y) = 0 \leftrightarrow x = y \qquad \textit{reflexividade}$$
$$(3)\ d(x,y) = d(y,x) \qquad \textit{simetria}$$
$$(4)\ d(x,y) \leq d(x,z) + d(y,z) \qquad \textit{desigualdade triangular}$$

A desigualdade triangular é considerada uma das propriedades mais importantes, pois é utilizada para estabelecer os limites do valor da distância entre dois objetos sem a necessidade do cálculo real da distância, acelerando os algoritmos de consulta por similaridade (FERREIRA et al., 2011). Considere os valores de distância d(x,z) e d(y,z), os limites para o valor de d(x,y) são $|d(x,z) - d(y,z)| \leq d(x,y) \leq d(x,z) + d(y,z)$.

Uma exemplificação de espaço métrico é ilustrada na Figura 5, que mostra os registros de dados representados por três classes de dados (vermelho, azul e verde), e um elemento de consulta representado pela estrela.

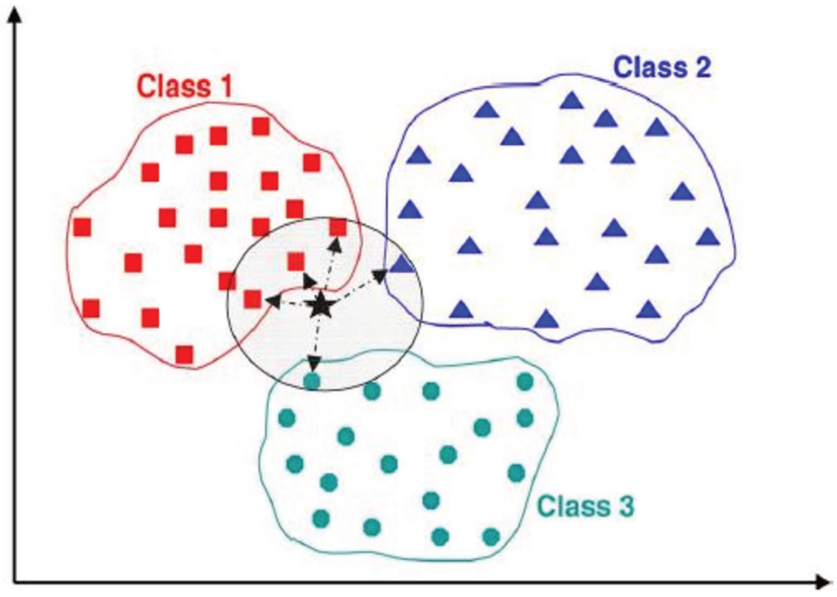

Figura 5: Espaço Métrico – Extraído de (CHAOVALITWONGSE et al., 2007)

2.3.1.2. Espaço Multidimensional

Se os objetos do domínio *S* correspondem a vetores de valores numéricos, o espaço é chamado Espaço Multidimensional ou Espaço Vetorial com Dimensão Finita (MACKUTE-VARONECKIENE et al., 2009). Os objetos de um espaço multidimensional de dimensão *n* (ou *n*-dimensional) são representados por *n* coordenadas de valores reais. Essa representação é ilustrada na Figura 6, em que o eixo x representa os dias, o eixo y mostra o índice de verde de uma determinada vegetação e na legenda o horário observado compreendido entre 6 e 18 horas.

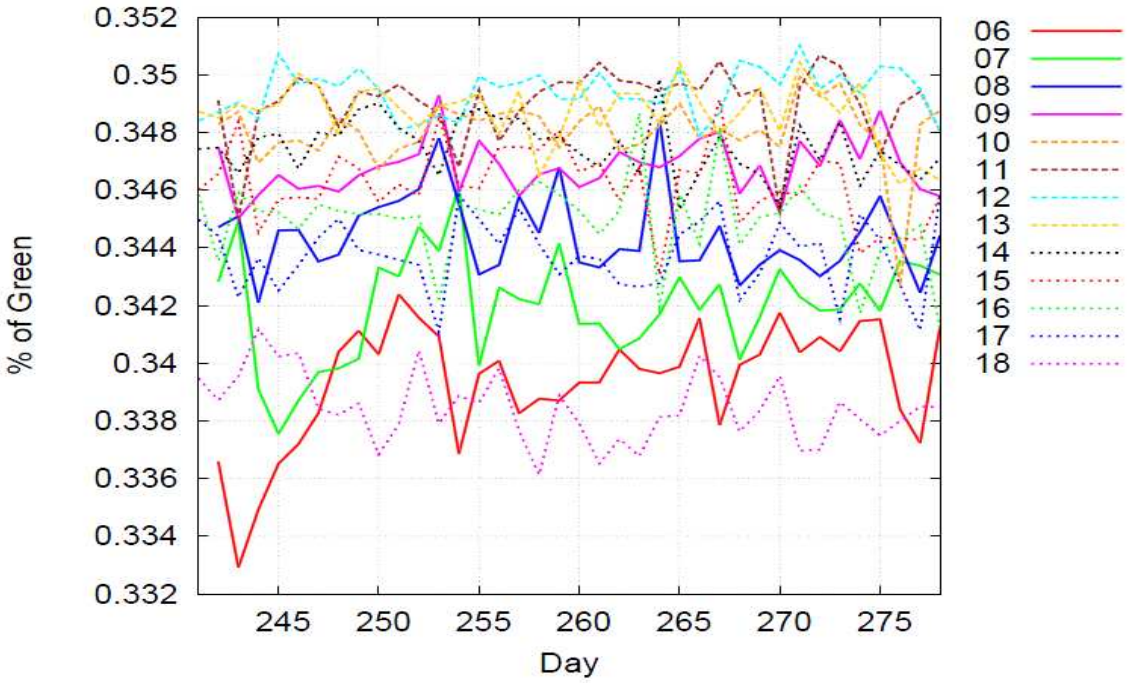

Figura 6: Espaço Multidimensional - extraída de: http://www.recod.ic.unicamp.br/ephenology/

2.3.2 Medidas de Distância

Existem várias medidas de distância para análise de séries temporais na literatura. Essas medidas são baseadas no cálculo de distâncias entre pares de objetos (séries temporais) de um conjunto de dados (ESLING & AGON, 2012). Nesse processo, é aplicada a função de distância nos valores dos atributos de interesse sobre os objetos (BATISTA et al., 2011).

Com o uso de uma determinada medida de distância, é possível descobrir a similaridade entre dois objetos, ou seja, podem-se identificar objetos que mais se aproximam, de acordo com suas características. Após descobrir os objetos com

características semelhantes, são identificados os vizinhos mais próximos através de algoritmos de mineração. Uma aplicação imediata consiste no uso de medidas de distância/similaridade em tarefas de agrupamento, ou seja, na identificação de objetos com características semelhantes entre si pertencentes ao mesmo grupo e diferentes em relação aos outros grupos. Outra aplicação que se beneficia da identificação de vizinhos mais próximos, consiste no uso da tarefa de classificação de séries temporais (ESLING & AGON, 2012).

As medidas de distância têm sido aplicadas pelos pesquisadores para facilitar o processamento de consultas e mineração de dados de séries temporais (MARIOTE et al., 2011; KEOGH, 2002). No entanto, com a grande variedade de técnicas propostas, verifica-se a necessidade de se comparar essas medidas de distância. Algumas dessas medidas são apresentadas a seguir:

2.3.2.1. Medida de distância – Manhattan

Entre as medidas mais conhecidas, encontra-se a distância de Manhattan, que consiste no cálculo baseado na soma das diferenças de uma determinada quantidade de valores de atributos (YI & FALOUTSOS, 2000). Por exemplo, considere dois objetos $a = (a_1, a_2,...,a_k)$ e $b = (b_1, b_2,...,b_k)$, a distância de Manhattan (também conhecida como L1) é obtida pela equação (1) a seguir:

$$L1(T_a, T_b) = \sum_{i=1}^{K} |T_a[i] - T_b[i]| \qquad (1)$$

Sendo que a e b são as séries temporais e k representa o número de atributos dessas séries.

Um exemplo de aplicação da distância de Manhattan é o trabalho de (MARIOTE et al., 2011), no qual é apresentado o descritor TIDES (*TIme series oscillation DEScriptor*). O TIDES descreve as séries temporais sob uma premissa diferente, a de caracterizar a oscilação dos dados e não seus valores (MARIOTE et al., 2011). Essa abordagem usa os coeficientes angulares de uma segmentação linear da curva que representa a evolução desses dados analisados em múltiplas escalas. Esse descritor permite a comparação e a mineração de séries utilizando várias granularidades, enriquecendo sua análise. O TIDES foi testado pelos autores em dois grupos de dados. O primeiro grupo chamado de treinamento e testes utilizando 600 séries e 60 pontos cada, o segundo grupo composto por dados reais de temperatura média mensal de 5 cidades

do Estado de São Paulo, totalizando 1336 séries de 48 pontos cada (MARIOTE et al., 2011).

Os resultados do descritor foram comparados por MARIOTE et al., (2011) com outro modelo, o *Linear Scan*, conforme apresentado na Tabela 1. As colunas TESTE1 e TESTE2 representam respectivamente o percentual de respostas corretas para o primeiro e segundo teste. Os resultados mostram que TIDES é mais apropriado em relação ao *Linear Scan*. No TESTE2, TIDES apresenta melhores resultados comparando com o *Linear Scan*, para um número pequeno de séries retornadas na consulta (K = 50), nos TESTE1 e TESTE2 TIDES representou melhores resultados considerando (K= 100). Somente quando o K passa a ser incremental ocorre que o fator de oscilação é aplicado mostrando sua influência, onde em ambos os testes TIDES apresenta melhor resultado (K = 100). Uma das vantagens em utilizar o descritor é justamente pelo fato que a análise está baseada na oscilação dos dados, não considerando os valores, com isso sua representação é imune ao eixo y. Porém, segundo os autores, esse descritor deve ser aplicado em outras bases de dados para comprovar sua eficácia.

Tabela 1: Comparação entre TIDES e *Linear Scan*

K	TESTE 1		TESTE 2	
	Linear Scan	TIDES	Linear Scan	TIDES
30	90	81(-9%)	78	77(-1%)
50	85	78(-7%)	73	74(1%)
70	78	76(-2%)	69	72(3%)
100	70	73(3%)	62	69(7%)

2.3.2.2. Medida de distância - Euclidiana

A distância Euclidiana é definida como uma função de distância entre dois objetos (FALOUTSOS et al., 1994). Seu cálculo baseia-se na soma das diferenças dos quadrados. Formalmente, para dois objetos a= (a_1,a_2,...,a_n) e b= (b_1,b_2,...,b_n) de mesmo tamanho K, define-se a distância L2 (a,b) como apresentado na equação (2):

$$L2(T_a, T_b) = \left(\sum_{i=1}^{K} (T_a[i] - T_b[i])^2\right)^{1/2} \quad (2)$$

2.3.2.3. Medida de distância – *Dynamic Time Warping*

A medida *Dynamic Time Warping (DTW),* segundo BERNDT & CLIFFORD, (1994), é mais tolerante a ruídos e informações faltantes. Essa medida faz o casamento das amostras de dois objetos a serem comparados de forma elástica, ou seja, uma série pode ser comparada com outra localizada em unidades de tempo antes ou depois dela. DTW é definida para dois objetos a e b com tamanho m e n como destacado nas equações (3) e (4):

$$DistDTW(T_a, T_b) = DTW(m, n) \tag{3}$$

sendo

$$DTW(i,j) = \begin{cases} 0 & se\ \ i = 0, j = 0 \\ \infty & se\ \ i = 0 \\ \infty & se\ \mathrm{j} = 0 \\ dist(T_a[i], T_b[j]]) + & \\ min\{DTW(i-1, j-1), & \\ DTW(i-1, j), DTW(i, j-1)\} & c.c. \end{cases} \tag{4}$$

Sendo que a medida DTW é calculada por programação dinâmica, ou seja, ela precisa encontrar como minimizar o valor da distância entre duas séries temporais (KEOGH, 2002). Para isso, preenche-se uma matriz N x M em que N e M são os tamanhos das séries temporais a serem comparadas. Basicamente, essa equação possui comportamentos específicos, que depende dos valores das variáveis de entrada, como por exemplo, a restrição de localidade na matriz.

KEOGH, (2002) apresenta uma nova técnica que possibilita um tempo de resposta mais rápido para a indexação da medida DTW. Nessa abordagem, primeiramente, aplica um limite no algoritmo de medida de distância para acelerar a busca sequencial considerando uma determinada consulta em seguida, adiciona uma técnica de indexação na medida DTW (KEOGH, 2002). A técnica de indexação proposta nesse trabalho é importante para responder consultas usando uma estrutura multidimensional.

Segundo o autor, esta técnica é atrativa devido a sua simplicidade, bem como, por ser intuitiva e competitiva em relação as outras abordagens mais complexas.

2.3.2.4. Medida de distância – *Longest Common Subsequence*

A medida de distância *Longest Common Subsequence* (LCSS) é tolerante a ruídos, ou seja, não sofre impactos devido às informações faltantes, e permite que uma série possa ser comparada com outra localizada em unidades de tempo não

necessariamente sequencial (VLACHOS et al., 2002). Essa medida é definida conforme apresentado nas equações (5) e (6):

$$LCSS(i,j) = \begin{cases} 0 & se\ \ i = 0 \\ 0 & se\ \ j = 0 \\ 1 + LCSS(i-1, j-1) & se\ \ |T_a[i] - T_b[i]| < \epsilon \\ max\{LCSS(i-1, j), LCSS(i, j-1)\} & c.c. \end{cases} \quad (5)$$

$$dist_{LCSS}(T_a, T_b) = 1 - \frac{LCSS(A,B)}{min(n,m)} \quad (6)$$

A medida LCSS permite escolher uma subsequência de valores da primeira e da segunda séries temporais que maximiza o número de valores considerados "iguais", ou seja, cuja diferença seja menor que um determinado valor ε (Epsilon), (VLACHOS et al., 2002.

Na abordagem proposta por VLACHOS et al., (2002), os autores investigam técnicas para análise e recuperação de objetos em um espaço de duas ou três dimensões. Nesse trabalho, os autores aplicam a medida de similaridade LCSS, pois além de ser tolerante a ruídos, isto é, não sofre impactos devido às informações faltantes, fornece uma noção intuitiva de semelhança entre as séries e permite alongamento de sequências no tempo.

A técnica empregada consiste no desenvolvimento de uma estrutura de indexação baseada em agrupamento hierárquico para responder as consultas de vizinhos mais próximos. Segundo os autores, os experimentos indicam que a técnica proposta pode ser usada para obter uma estimativa precisa e rápida da distância entre dois objetos mesmo na presença de ruídos.

2.3.2.5. Medida de distância – *Zero-mean Normalized Cross Correlation*

Outra medida de distância utilizada na literatura é a *Zero-mean Normalized Cross Correlation* (ZNCC) -- correlação cruzada normalizada com média zero. Esta medida é um método padrão para estimar o grau de correlação entre duas séries temporais (MARTIN & CROWLEY, 1995). Os dados de entrada são normalizados, os coeficientes de correlação podem variar de -1 a 1, indicando os limites máximos de correlação, e 0 que indica que não há correlação A medida está definida para dois objetos ou sinais a e b conforme apresentado na equação (7):

$$ZNCC(T_a, T_b) = \max_x \frac{\sum_{i=0}^{n-1}(T_a[i] - \overline{T_a})(T_b[x+i] - \overline{T_b})}{\sqrt{\sum_{i=0}^{n-1}(T_a[i] - \overline{T_a})^2 \times \sum_{i=0}^{m-1}(T_b[x+i] - \overline{T_b})^2}}$$

$$onde \overline{T_a} e \overline{T_b} \; sao \; as \; medias \; dos \; valores \; T_a \; e \; T_b, respectivamente. \quad (7)$$

2.3.2.6. Medida de distância – *Edit distance with Real Penalty*

CHEN et al., (2004), a abordagem proposta é baseada na medida *Edit distance with Real Penalty (ERP*). Essa medida é considerada uma nova função de distância, pois é composta por características de medidas de similaridade tais como L2 e DTW (CHEN et al., 2004), e está definida na equação (8):

$$ERP(T_a, T_b) = \begin{cases} \sum_{i=1}^{n} dist(T_b[i], g) & se \;\; m = 0 \\ \sum_{i=1}^{m} dist(T_a[i], g) & se \;\; n = 0 \\ min\{ERP(Rest(T_a), Rest(T_b)) + dist(T_c[1], T_b[1]), & \\ ERP(Rest(T_a), T_b) + dist(T_a[1], g), & c.c. \\ ERP(T_a, Rest(T_b)) + dist(T_b[1], g)\} & \end{cases} \quad (8)$$

Sendo que A e B são duas séries temporais e g é um valor constante, ou seja, indica o valor de g, em caso de gap entre as séries temporais. Para quaisquer objetos, nos quais qualquer um deles sendo um objeto de intervalo, é necessário que a distância seja: *dist (ai,ci) ≤ dist (ai,bi) + (bi,ci)*

2.3.2.7. Medida de distância – *Edit Distance on Real sequence*

A medida *Edit Distance on Real sequence (EDR*) também foi proposta por CHEN et al., (2005) para resolver alguns problemas encontrados em outras medidas de similaridade, como por exemplo a baixa acurácia devido a intervalos entre séries temporais. Dado dois objetos A e B de tamanhos n e m, respectivamente, a distância EDR entre A e B é o número de inclusão, exclusão ou substituição de operações que são necessárias para alterar A em B. A medida EDR (A,B) é definida na equação (9):

$$EDR(T_a, T_b) = \begin{cases} n & se \;\; m = 0 \\ m & se \;\; n = 0 \\ min\{EDR(Rest(T_a), Rest(T_b)) + subcost, & \\ EDR(Rest(T_a), T_b) + 1, & \\ EDR(T_a, Rest(T_b)) + 1\} & c.c. \end{cases} \quad (9)$$

Sendo que subcost = 0, se igualdade (*Ta*[1],*Tb*[1]) = verdadeiro e subcost = 1, caso contrário e Rest(T) representa para a série temporal obtida a partir de T eliminando o último elemento.

Nesse trabalho, foram desenvolvidas algumas técnicas para melhorar o resultado de buscas e cálculos de distâncias nos valores dos atributos em bancos de dados. Segundo os autores, essa técnica pode ser mais eficiente em relação as já existentes, uma vez que não é sensível a ruídos, bem como não sofre impactos quanto a pares de objetos de diferentes trajetórias.

2.3.3 Comparações entre Medidas de Similaridade

Na comparação entre as medidas de similaridade apresentadas anteriormente, destaca-se o trabalho de DING et al., (2008). Nessa análise, os autores citam por exemplo a distância Euclidiana como uma medida de similaridade simples, uma vez que essa medida é suficiente para cálculo de similaridade em diversas bases de dados.

Por outro lado, segundo DING et al., (2008), a medida DTW (*Dynamic Time Warping*) realiza alinhamento entre duas séries, reconhecendo sua similaridade mesmo ocorrendo deslocamentos de tempo. Os autores destacam ainda, a importância do uso de outras bases de dados para melhorar a validação e comparação entre as medidas de similaridade.

Outro trabalho de pesquisa que analisa e compara algumas medidas de similaridade é o estudo de ESLING & AGON, (2012). Nesse trabalho, os autores citam que a escolha de uma medida de similaridade adequada depende da natureza do dado a ser analisado assim como a aplicação específica para essa finalidade. Se a análise de séries temporais se concentra em uma base relativamente pequena e a percepção visual for significativa, uma medida baseada em forma pode ser mais apropriada (ESLING & AGON, 2012). Se a aplicação tem como alvo um conjunto de dados muito específicos ou qualquer tipo de conhecimento prévio sobre esses dados que estão disponíveis, métodos baseados em modelos estatísticos podem fornecer uma abstração mais importante do ponto de vista de interesse do usuário (ESLING & AGON, 2012).

Métodos baseados em características podem ser mais relevantes quando a periodicidade do tempo for a variável mais interessante. Finalmente, quando a série temporal se tratar de grandes volumes de dados, um pequeno conhecimento sobre a estrutura é necessário. De qualquer modo, mesmo com as recomendações e comparações para a seleção mais apropriada de medidas de similaridade, a acurácia da similaridade escolhida também deve ser avaliada (ESLING & AGON, 2012). Porém,

essa não é uma tarefa muito simples, sendo que a acurácia de uma medida de similaridade seja normalmente avaliada, usando o algoritmo *K-Nearest Neighbor (KNN)*.

ESLING & AGON, (2012) ressaltam ainda algumas particularidades dessas medidas de distância, classificando-as como, por exemplo, em medidas baseadas em forma, edição e características. Medidas baseadas em forma incluem a distância Euclidiana e a DTW, sendo que o tempo de processamento da medida Euclidiana é melhor em relação a DTW. Medidas baseadas em edição incluem a medida ERP e LCSS, sendo que o tempo de processamento da medida ERP é mais custoso em relação ao tempo de processamento da medida LCSS, ou seja, a complexidade da medida ERP é $O(n^2)$, por outro lado, a medida LCSS tem complexidade $O(n)$. Medidas baseadas em características incluem a autocorrelação e histograma, sendo que o tempo de processamento do histograma é menos custoso em relação a autocorrelação.

A Tabela 2 apresenta algumas características dessas medidas de distância pesquisadas.

Tabela 2: Principais características das medidas de distância

Medida	Referência	Vantagens	Desvantagens	Complexidade	Características
Distância Euclidiana (L2)	(FALOUTSOS et al., 1994) (ESLING & AGON, 2012)	Avaliação linear, fácil de ser implementada	Sensível a ruídos	O(n)	Baseada em forma
Distância de Manhattan (L1)	(YI & FALOUTSOS, 2000) (ESLING & AGON, 2012)	Cálculo simples	Nem sempre produz a menor distância de um ponto a outro	O(n)	Baseada em forma
Dynamic Time Warping (DTW)	(KEOGH, 2002) (ESLING & AGON, 2012)	Tolerância a ruídos e informações faltantes	Tempo de processamento alto	$O(n^2)$	Baseada em forma
Longest Common Subsequence (LCSS)	(VLACHOS et al., 2002) (ESLING & AGON, 2012)	Tempo de processamento baixo	Acurácia baixa na presença de ruídos	O(n)	Baseada em edição
Zero-mean Normalized Cross Correlation (ZNCC)	(MARTIN & CROWLEY, 1995)	Dados normalizados	Tempo de processamento alto	O(n)	Baseada em características
Edit distance with Real Penalty (ERP)	(CHEN et al., 2004) (ESLING & AGON, 2012)	Eficiente para indexação	Sensível a ruídos	$O(n^2)$	Baseada em edição
Edit Distance on Real sequence (EDR)	(CHEN et al., 2005) (ESLING & AGON, 2012)	Tolerância a ruídos	Tempo de processamento alto	$O(n^2)$	Baseada em edição

Esse trabalho selecionou um conjunto de medidas de distância para avaliação de sua eficácia, ou seja, analisou acertos e erros dessas medidas do ponto de vista de classificação de séries temporais. Além disso, não considerou as vantagens, desvantagens e complexidade algorítmica para calcular as medidas de similaridade.

2.4 Tarefas de Mineração de Dados

Existem várias abordagens na literatura relacionadas com as tarefas de mineração de dados em séries temporais. Estas tarefas são usadas para encontrar padrões dos dados a partir das etapas utilizadas no processo de descoberta do conhecimento em bancos de dados (WEI, 2009).

As tarefas basicamente são divididas em dois objetivos, sendo previsão, que envolve o uso de atributos no banco de dados para prever valores futuros de outros atributos de interesse e descrição, que busca encontrar padrões de interesse do usuário com base na descrição dos dados (FAYYAD et al., 1996).

Os objetivos da previsão e descrição podem ser alcançados usando uma variedade de tarefas ou métodos específicos de mineração de dados. Essas tarefas são propostas como classificação, regressão, agrupamento, correlação e sumarização (FAYYAD et al., 1996).

A seguir são apresentados alguns conceitos e abordagens relacionadas à tarefa de Classificação, bem como o algoritmo *k-Nearest Neighbor (KNN*), (CHAOVALITWONGSE et al., 2007; HAN et al., 2001). Esta tarefa será utilizada para avaliar as medidas de distância, que foram implementadas no contexto do problema de classificação de séries temporais, considerando a similaridade dos registros com os vetores de características gerados.

2.4.1 Classificação

A tarefa de mineração de dados de séries temporais baseada em classificação tem despertado grande interesse nos últimos anos (CHAOVALITWONGSE et al., 2007; HAN et al., 2001) .

FAYYAD et al., (1996), define classificação como sendo uma tarefa que tem como objetivo identificar e avaliar uma determinada função para classificar um dado em diversas classes pré-definidas. Essa tarefa possui a função de classificar métodos como parte da descoberta do conhecimento em várias aplicações e para identificar automaticamente objetos de interesse em Bancos de Dados.

Segundo HAN et al., (2001) a classificação consiste em examinar uma característica nos dados e atribuir uma classe previamente definida. Nessa abordagem, os dados são associados a classes ou a conceitos utilizando-se de um processo de discriminação ou de caracterização.

Discriminação tem seu resultado obtido a partir da atribuição de um valor a um atributo. Caracterização é a sumarização de um atributo de interesse por uma característica de um ou mais atributos (HAN et al., 2001).

Nas abordagens relacionadas à tarefa de classificação, o algoritmo dos vizinhos mais próximos é usado no sentido de classificar os dados associados a essa tarefa (KASHYAP et al., 2011). Na abordagem proposta por KASHYAP et al., (2011), uma consulta utilizando o algoritmo K-*Nearest Neighbor* (K-NN) obtém um resultado para os K registros mais similares em uma base de dados, considerando uma consulta pré-definida, baseado em uma medida de similaridade.

Para CHAOVALITWONGSE et al., (2007), o K-NN retorna os K elementos mais similares dado um determinado elemento representante. Neste trabalho, o classificador rotula objetos com base na sua similaridade entre amostras dos dados de treinamento. Para dados não rotulados de séries temporais, a regra do K-NN encontra os K vizinhos mais próximos de séries temporais rotuladas no conjunto de dados de treinamento e atribui um valor para a classe que aparece com mais frequência na vizinhança das séries rotuladas. Além dos dados de treinamento, a regra do K-NN requer dois parâmetros de entrada que são usados para a classificação de uma nova série temporal não rotulada. Esses parâmetros basicamente são: o tamanho dos K vizinhos mais próximos e uma função de distância, que é usado como uma medida de proximidade.

Ainda em CHAOVALITWONGSE et al., (2007), há duas regras comuns para a classificação de novos dados não rotulados: o voto por maioria e por similaridade. Em votação por maioria, uma classe recebe uma votação para cada instância da classe no conjunto da vizinhança de K das amostras. Em seguida, as amostras de novos dados são classificadas na classe com a maior quantidade de votos. Na soma da pontuação de similaridade, cada classe recebe uma pontuação que é igual a soma dos valores de similaridade das instâncias da classe do conjunto de amostras da vizinhança de K. Em seguida, a amostra de novos dados é classificada para a classe com o maior valor da soma de similaridade.

2.5 Considerações Finais

No processo de mineração de dados, a análise de séries temporais se tornou uma importante fonte de investigação em várias áreas, despertando grande interesse de pesquisadores particularmente para a avaliação de medidas de similaridade para análise de séries temporais. Essas medidas de similaridade são usadas para calcular as distâncias entre objetos de um conjunto de dados, para os quais é aplicada uma função de distância nos valores de interesse. Entre as medidas de distância destacam-se L1 (*Manhattan*), a distância L2 (Euclidiana), a DTW, a LCSS, ZNCC, ERP e EDR.

Essas e outras medidas de distância são importantes para a tarefa de classificação, bem como o algoritmo K-*Nearest Neighbor* (K-NN). Esta tarefa será utilizada para avaliar a eficácia das medidas de similaridade no contexto do problema de classificação de séries temporais no domínio de Fenologia, considerando as características extraídas de imagens de vegetação.

3 Fenologia

Este capítulo apresenta conceitos relacionados à fenologia e algumas pesquisas sobre o assunto.

3.1 Considerações Iniciais

O estudo de fenologia tem despertado interesse da comunidade científica como uma área que possibilita entender o efeito das alterações climáticas em espécies (CASTRO et al., 2007). Exemplos de estudos incluem análise florestal e ecologia, pois investiga os fenômenos naturais repetitivos, como os processos de floração, frutificação e mudança foliar e sua relação com o clima (ALBERTI et al., 2008). Muitos desses processos, como a queda de folhas e a floração, estão normalmente relacionados com as mudanças climáticas (PINTO et al., 2000).

3.2 Pesquisas em Fenologia

Fenologia é definida como o estudo de fenômenos naturais recorrentes e sua relação com o clima (SCHWARTZ, 2003). É uma ciência que observa os ciclos de plantas e animais e sua relação com os dados meteorológicos locais (RATHCKE & LACEY, 1985).

As pesquisas recentes na área de fenologia buscam mostrar a importância da análise de padrões em séries temporais associadas a fenômenos fenológicos (CASTRO et al., 2007; ALBERTI et al., 2008). Essas séries temporais são registradas por sensores ou radares, como por exemplo, o sensor MODIS *(Moderate Resolution Imaging Spectroradiometer)* EVI (*Enhanced Vegetation Index*) e NDVI (*Normalized Difference Vegetation Index*), (ZHANG et al., 2009; COUTO et al., 2011).

Além das abordagens recentes publicadas, MORELLATO et al., (2013) apresenta uma revisão em pesquisas em fenologia na América Central e América do Sul. Nesse trabalho, são descritos os padrões de floração e frutificação de principais tipos de vegetação, e destacadas áreas onde faltam informações fenológicas. Pesquisas em fenologia são ainda necessárias, mesmo com o aumento de publicações disponíveis, principalmente no século XXI. Floresta tropical úmida continua a ser, de longe, o ecossistema mais amplamente estudado, enquanto árvores são a forma de vida observada em quase todos os jornais pesquisados (MORELLATO et al., 2013).

Atualmente, os conjuntos de dados fenológicos a longo prazo são raros e alguns sistemas de monitoramento de longo prazo são conhecidos na América do Sul e

Central. Poucos artigos revisados estavam preocupados com os efeitos da mudança climática e sua avaliação usando fenologia de planta, o que difere muito de pesquisas no Hemisfério Norte, que teve uma forte concentração na fenologia e nas mudanças climáticas. Dentre as novidades podemos citar fenologia e efeitos de borda e fragmentação, e monitoramento remoto de fenologia com câmeras digitais. Por fim, a construção de redes em fenologia é o maior desafio para pesquisadores na América do Sul e da América Central, exigindo um esforço de cooperação entre Universidades e, Instituições de pesquisa e agências governamentais e não-governamentais (MORELLATO et al., 2013).

Em ZHAO et al., (2012) analisam as imagens de três espécies (duas de uma floresta de folhas verdes e outra de uma floresta tropical sazonal). Essas imagens foram utilizadas para estimar os eventos fenológicos de desenvolvimento das folhas e seu envelhecimento.

Em CHAMBERS et al., (2013) discutem como informações fenológicas podem informar a capacidade de adaptação das espécies, sua capacidade de resistência, e as restrições de adaptação autônoma. Além disso, destacam também deficiências graves na coleta de dados tanto no passado como no presente para análises em grandes escalas regionais (com poucos estudos nos trópicos e na África). Adicionalmente, discutem se as previsões precisas sobre os efeitos gerais das mudanças climáticas sobre a biologia de organismos são realizadas e deficiências nas políticas de coleta de dados centrados na segmentação das regiões e taxa de dados são observadas.

3.3 Considerações Finais

Os pesquisadores devem buscar ferramentas e técnicas que os auxiliem a monitorar a fenologia de plantas e animais e compreender a sua relação com o clima. Além disso, com o auxílio da tecnologia criam-se condições favoráveis para armazenar e analisar esse volume de dados que são coletados diariamente por sensores e radares.

4 Desenvolvimento

Este capítulo apresenta o detalhamento dos procedimentos adotados durante o desenvolvimento desta pesquisa e as experimentações selecionadas para a avaliação de resultados que atendam ao objetivo deste trabalho.

4.1 Organização dos Dados

Os dados usados no processo de avaliação da acurácia das medidas estão separados em dois grupos:

- Séries de Testes : é composto por regiões contendo indivíduos de três espécies. Foram selecionados considerando as espécies *Aspidosperma tomentosum, Miconia rubiginosa* e *Pouteria ramiflora*.
- Séries de Treino : é composto por regiões contendo indivíduos de seis espécies. Foram selecionados considerando as espécies *Aspidosperma tomentosum*, *Caryocar brasiliensis, Myrcia guianesis, Miconia rubiginosa, Pouteria torta* e *Pouteria ramiflora*.
  - Rotulados : corresponde a 398 regiões.
  - Não Rotulados: corresponde a 8415 regiões de espécies desconhecidas.

Para a realização das avaliações, estas séries foram organizadas nas seguintes classes:

- Espécies: *Aspidosperma tomentosum, Miconia rubiginosa, Pouteria ramiflora, Caryocar brasiliensis, Myrcia guianesis* e *Pouteria torta*
- Canais de Cor: R, G e B (*Red, Green e Blue*)
- Hora: das 6 às 18 horas.

4.2 Procedimentos de Avaliação

Os resultados foram verificados e validados seguindo os procedimentos e critérios abaixo relacionados.

Os procedimentos realizados para a execução dos testes ocorreram da seguinte forma:

- Cálculo das distâncias de uma série de teste para todas as séries do conjunto de treino (séries com rótulos conhecidos).
- Ordenação dessas séries, e uso do classificador K-NN.
- Cálculo dos acertos e erros do classificador para os conjuntos de séries de teste.

Os experimentos usados nesse trabalho consistiram em comparar a eficácia das medidas de similaridade considerando as visões de análise geral por hora, por canal de cor e por espécie, considerando o primeiro vizinho mais próximo.

Na Tabela 3, é ilustrado um esquema proposto de como será efetuada a validação dos resultados. Nessa tabela é apresentado o horário de análise, o canal de cor, o ID da região de teste, ID da classe de teste, Nome da classe de teste, ID Região 1-NN, ou seja, o vizinho mais próximo, ID da classe 1-NN, Nome da classe 1-NN, a medida de dsitância utilizada, e o resultado que indica se o classificador acertou ou errou.

Tabela 3: Exemplo de dados para validação

Acurácia do Classificador

Hora: 6 Canal de cor: ***Green***

ID Região Teste	Id da Classe de Teste	Nome da Classe de Teste	ID Região 1-NN	ID da Classe 1-NN	Nome da Classe 1-NN	Medida de Distância	Resultado
2390	6	*Aspidosperma tomentosum*	1107	0	*Aspidosperma tomentosum*	L1	Acerto
2390	6	*Aspidosperma tomentosum*	0	0	Desconhecida	L1	Erro
4205	6	*Aspidosperma tomentosum*	4629	1	*Caryocar brasiliensis*	L2	Erro
3407	7	*Miconia rubiginosa*	2746	3	*Miconia rubiginosa*	L2	Acerto
3407	7	*Miconia rubiginosa*	4645	2	*Myrcia guianesis*	DTW	Erro
4178	8	*Pouteria ramiflora*	2746	3	*Miconia rubiginosa*	DTW	Erro
2453	8	*Pouteria ramiflora*	3462	3	*Miconia rubiginosa*	LCSS	Erro
2390	6	*Aspidosperma tomentosum*	0	0	Desconhecida	LCSS	Erro
4205	6	*Aspidosperma tomentosum*	2746	3	*Miconia rubiginosa*	ZNCC	Erro
3407	7	*Miconia rubiginosa*	2746	3	*Miconia rubiginosa*	ZNCC	Acerto
2390	6	*Aspidosperma tomentosum*	1107	0	*Aspidosperma tomentosum*	EDR	Acerto
4178	8	*Pouteria ramiflora*	4645	2	*Myrcia guianesis*	EDR	Erro
4205	6	*Aspidosperma tomentosum*	1107	0	*Aspidosperma tomentosum*	ERP	Acerto
3407	7	*Miconia rubiginosa*	0	0	Desconhecida	ERP	Erro

O critério utilizado para definição do resultado considerou que se o vizinho mais próximo da espécie da classe de treino pertencer a mesma espécie da classe de teste, nesse caso o classificador, baseado em uma medida de distância acertou, caso contrário, o classificador errou.

4.3 Análise dos Resultados

Nessa seção são apresentadas algumas avaliações sobre as medidas de distância que revelam conhecimentos relevantes para especialistas na área de fenologia e mineração de dados.

4.3.1 Análise considerando horários

Na visão geral por horário, identificou-se na Tabela 4 que a medida de distância ERP apresentou o melhor percentual de acerto com 29,90% e que as 17 horas apresentou o melhor percentual de acerto com 43,24%.

Tabela 4: Análise geral por horário

Hora	L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
6	35,81%	38,96%	31,08%	38,51%	33,78%	35,59%	39,41%	36,16%	2,83%
7	29,05%	31,08%	30,18%	27,25%	34,01%	36,26%	33,11%	31,56%	2,87%
8	23,20%	23,42%	21,62%	24,77%	24,10%	25,23%	24,77%	23,87%	1,15%
9	26,58%	25,45%	19,82%	25,23%	23,20%	22,75%	27,25%	24,32%	2,38%
10	23,20%	23,87%	13,96%	25,23%	27,70%	25,90%	25,45%	23,62%	4,16%
11	26,13%	24,77%	22,52%	26,58%	25,68%	28,38%	25,00%	25,58%	1,67%
12	22,97%	26,80%	29,05%	26,35%	28,38%	26,80%	27,93%	26,90%	1,84%
13	26,13%	26,13%	24,55%	27,03%	28,38%	28,15%	25,68%	26,58%	1,27%
14	25,23%	24,55%	17,34%	23,42%	26,35%	24,10%	24,77%	23,68%	2,72%
15	24,77%	25,45%	13,74%	26,80%	23,87%	23,65%	26,80%	23,58%	4,19%
16	24,77%	25,00%	20,50%	23,87%	24,55%	22,30%	24,55%	23,65%	1,54%
17	46,85%	44,82%	33,78%	45,72%	40,09%	44,82%	46,62%	43,24%	4,39%
18	35,14%	38,06%	21,17%	38,06%	36,94%	36,04%	37,39%	34,68%	5,60%
Média	**28,45%**	**29,11%**	**23,02%**	**29,14%**	**29,00%**	**29,23%**	**29,90%**		
Desvio Padrão	**6,65%**	**6,71%**	**6,16%**	**6,68%**	**5,25%**	**6,57%**	**6,79%**		

Adicionalmente, nesta mesma visão de análise, observou-se que a medida LCSS apresentou o melhor desvio padrão com 5,25% (Figura 7).

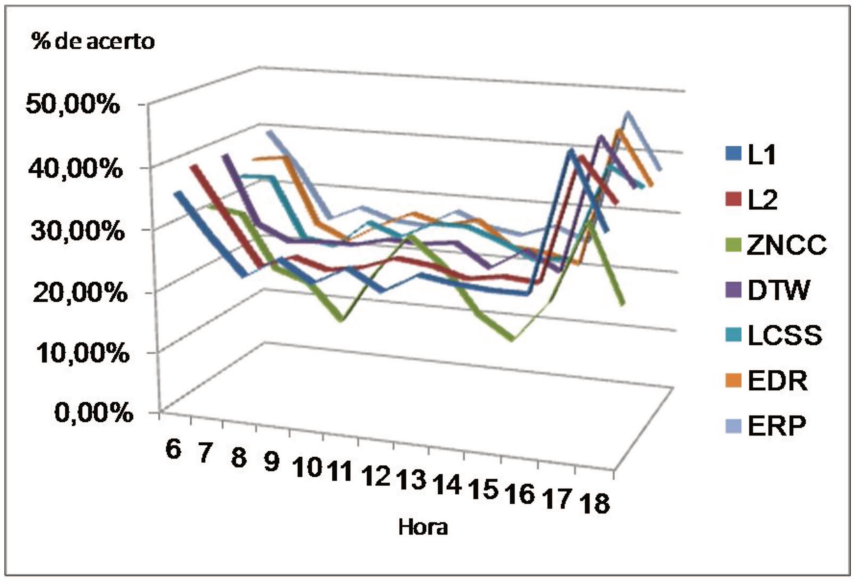

Figura 7: % Acerto 1-NN - Hora

4.3.2 Análise considerando canais de cor

Canal de cor: *Red*

Na visão por canal de cor *Red,* identificou-se na Tabela 5 que a medida de distância DTW apresentou o melhor percentual de acerto com 25,05% e que as 17 horas apresentou o melhor percentual de acerto com 53,09%.

Tabela 5: Análise por Canal de Cor - *Red*

Hora	L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
6	27,70%	32,43%	23,65%	34,46%	27,03%	26,35%	32,43%	29,15%	3,67%
7	21,62%	22,97%	29,05%	19,59%	33,78%	33,11%	22,97%	26,16%	5,33%
8	12,84%	17,57%	25,68%	25,68%	11,49%	17,57%	19,59%	18,63%	5,17%
9	20,95%	18,92%	18,24%	20,27%	15,54%	16,89%	22,30%	19,02%	2,18%
10	22,30%	20,95%	11,49%	25,00%	20,95%	18,24%	23,65%	20,37%	4,14%
11	22,30%	18,24%	24,32%	25,00%	16,89%	20,27%	16,89%	20,56%	3,15%
12	14,86%	16,89%	25,68%	18,92%	18,24%	16,89%	19,59%	18,73%	3,18%
13	12,16%	11,49%	20,27%	16,22%	14,19%	16,22%	11,49%	14,58%	3,00%
14	19,59%	18,92%	14,19%	12,84%	18,24%	17,57%	18,92%	17,18%	2,42%
15	20,95%	19,59%	12,84%	25,68%	16,89%	16,89%	22,30%	19,31%	3,89%
16	18,92%	18,92%	17,57%	17,57%	14,86%	13,51%	17,57%	16,99%	1,89%
17	58,78%	57,43%	42,57%	51,35%	47,30%	55,41%	58,78%	53,09%	5,81%
18	32,43%	34,46%	9,46%	33,11%	37,84%	36,49%	32,43%	30,89%	8,95%
Média	**23,49%**	**23,75%**	**21,15%**	**25,05%**	**22,56%**	**23,49%**	**24,53%**		
Desvio Padrão	**11,49%**	**11,39%**	**8,54%**	**9,68%**	**10,36%**	**11,37%**	**11,33%**		

Nesta mesma visão de análise, observou-se que a medida ZNCC apresentou o melhor desvio padrão com 8,54% (Figura 8).

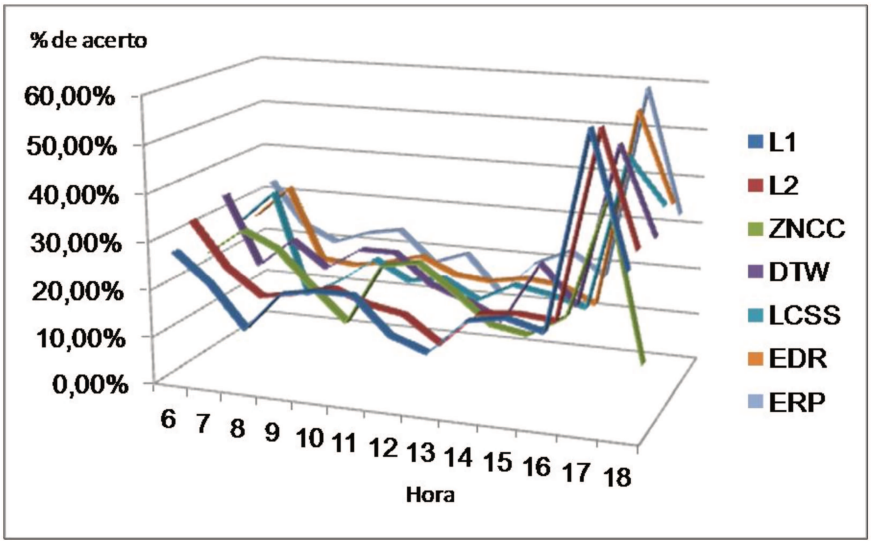

Figura 8: % Acerto 1-NN - Cor: *Red*

Canal de cor: *Green*

Na visão por canal de cor *Green,* identificou-se na Tabela 6 que a medida de distância LCSS apresentou o melhor percentual de acerto com 34,62% e que as 06 horas apresentou o melhor percentual de acerto com 40,35%.

Tabela 6: Análise por Canal de Cor - *Green*

Hora	L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
6	40,54%	42,57%	37,16%	39,86%	38,51%	40,54%	43,24%	40,35%	1,97%
7	35,14%	36,49%	38,51%	35,81%	35,81%	36,49%	37,84%	36,58%	1,11%
8	31,76%	32,43%	23,65%	27,03%	38,51%	34,46%	31,08%	31,27%	4,48%
9	31,76%	29,73%	18,92%	27,70%	29,05%	27,03%	29,05%	27,61%	3,81%
10	25,00%	23,65%	10,81%	27,03%	32,43%	31,76%	24,32%	25,00%	6,64%
11	26,35%	26,35%	19,59%	27,03%	31,08%	33,11%	27,03%	27,22%	3,95%
12	26,35%	30,41%	28,38%	27,03%	32,43%	31,08%	30,41%	29,44%	2,07%
13	29,73%	31,08%	22,97%	29,73%	35,81%	32,43%	29,73%	30,21%	3,59%
14	27,70%	27,03%	16,22%	28,38%	32,43%	32,43%	27,03%	27,32%	5,03%
15	33,11%	33,11%	16,22%	29,05%	33,11%	33,11%	35,81%	30,50%	6,11%
16	32,43%	31,08%	24,32%	29,05%	35,81%	33,11%	30,41%	30,89%	3,34%
17	37,16%	33,78%	33,78%	41,22%	35,81%	37,16%	33,78%	36,10%	2,52%
18	39,19%	39,86%	31,76%	43,24%	39,19%	39,19%	39,86%	38,90%	3,21%
Média	**32,02%**	**32,12%**	**24,79%**	**31,70%**	**34,62%**	**33,99%**	**32,28%**		
Desvio Padrão	**4,80%**	**5,07%**	**8,29%**	**5,82%**	**3,00%**	**3,46%**	**5,31%**		

Nesta mesma visão de análise, observou-se que a medida LCSS apresentou o melhor desvio padrão com 3,00% (Figura 9).

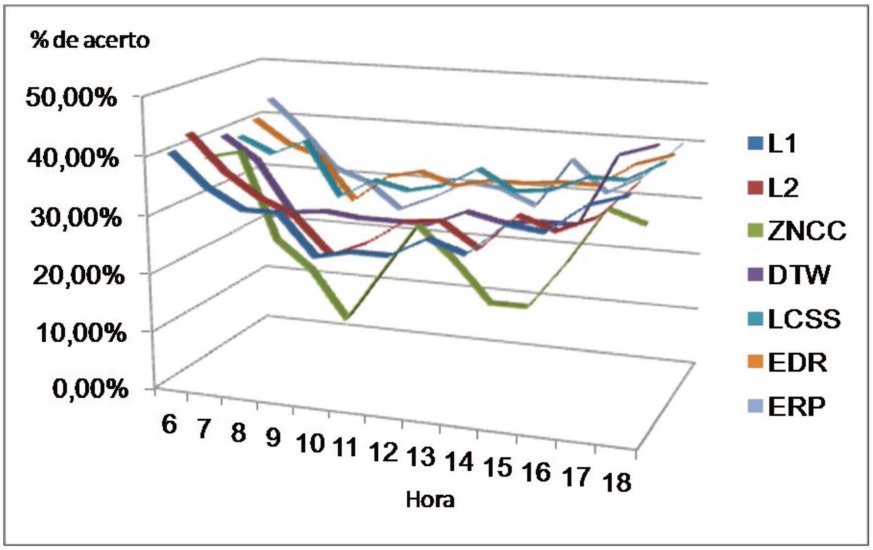

Figura 9: % Acerto 1-NN - Cor: *Green*

Canal de cor: *Blue*

Na visão por canal de cor *Blue,* identificou-se na Tabela 7 que a medida de distância ERP apresentou o melhor percentual de acerto com 32,90% e que as 17 horas apresentou o melhor percentual de acerto com 40,54%.

Tabela 7: Análise por Canal de Cor - *Blue*

Hora	L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
6	39,19%	41,89%	32,43%	41,22%	35,81%	39,86%	42,57%	39,00%	3,38%
7	30,41%	33,78%	22,97%	26,35%	32,43%	39,19%	38,51%	31,95%	5,52%
8	25,00%	20,27%	15,54%	21,62%	22,30%	23,65%	23,65%	21,72%	2,90%
9	27,03%	27,70%	22,30%	27,70%	25,00%	24,32%	30,41%	26,35%	2,48%
10	22,30%	27,03%	19,59%	23,65%	29,73%	27,70%	28,38%	25,48%	3,42%
11	29,73%	29,73%	23,65%	27,70%	29,05%	31,76%	31,08%	28,96%	2,49%
12	27,70%	33,11%	33,11%	33,11%	34,46%	32,43%	33,78%	32,53%	2,06%
13	36,49%	35,81%	30,41%	35,14%	35,14%	35,81%	35,81%	34,94%	1,90%
14	28,38%	27,70%	21,62%	29,05%	28,38%	22,30%	28,38%	26,54%	2,93%
15	20,27%	23,65%	12,16%	25,68%	21,62%	20,95%	22,30%	20,95%	3,96%
16	22,97%	25,00%	19,59%	25,00%	22,97%	20,27%	25,68%	23,07%	2,21%
17	44,59%	43,24%	25,00%	44,59%	37,16%	41,89%	47,30%	40,54%	6,98%
18	33,78%	39,86%	22,30%	37,84%	33,78%	32,43%	39,86%	34,27%	5,64%
Média	**29,83%**	**31,44%**	**23,13%**	**30,67%**	**29,83%**	**30,20%**	**32,90%**		
Desvio Padrão	**6,79%**	**6,93%**	**5,88%**	**6,85%**	**5,25%**	**7,26%**	**7,29%**		

Nesta mesma visão de análise, observou-se que a medida LCSS apresentou o melhor desvio padrão com 5,25% (Figura 10).

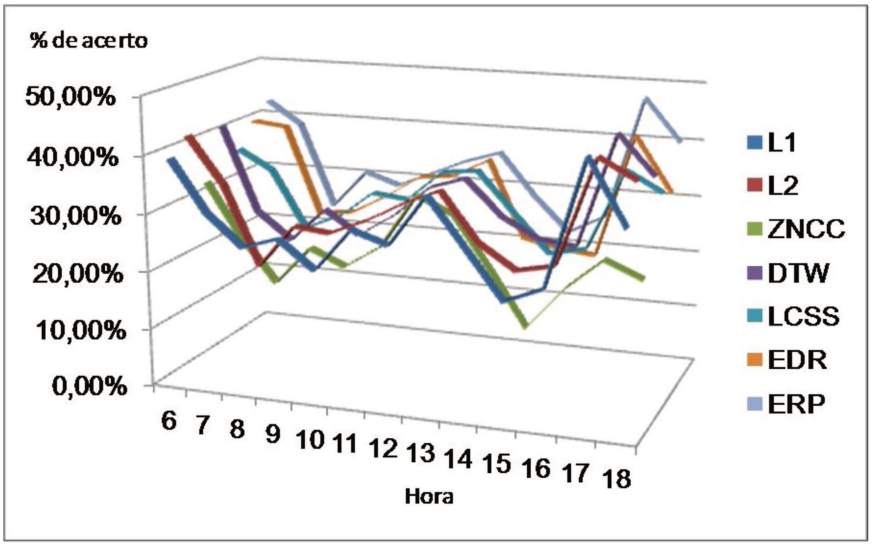

Figura 10: % Acerto 1-NN - Cor: *Blue*

4.3.3 Análise considerando Espécies

Espécie: *Aspidosperma tomentosum*

Na visão por espécie *Aspidosperma tomentosum,* identificou-se na Tabela 8 que a medida de distância LCSS apresentou o melhor percentual de acerto com 51,77% e que as 17 horas apresentou o melhor percentual de acerto com 77,63%.

Tabela 8: Análise por Espécie – *Aspidosperma tomentosum*

Hora	L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
6	66,67%	71,96%	58,73%	69,31%	66,14%	67,20%	71,43%	67,35%	4,11%
7	41,27%	46,03%	30,69%	37,57%	61,38%	67,20%	53,97%	48,30%	12,19%
8	37,57%	37,57%	30,69%	36,51%	47,09%	44,44%	40,74%	39,23%	5,03%
9	40,74%	38,10%	31,75%	38,10%	42,86%	39,68%	41,80%	39,00%	3,39%
10	33,33%	35,98%	22,22%	37,57%	47,09%	42,86%	38,10%	36,73%	7,27%
11	42,86%	41,80%	37,04%	41,80%	49,21%	51,32%	43,92%	43,99%	4,48%
12	37,04%	42,86%	42,33%	40,74%	50,79%	47,09%	43,92%	43,54%	4,10%
13	44,44%	43,92%	35,45%	46,56%	52,91%	49,74%	42,33%	45,05%	5,17%
14	34,92%	33,86%	24,34%	37,04%	42,33%	38,62%	34,39%	35,07%	5,16%
15	34,92%	36,51%	18,52%	37,04%	40,21%	38,10%	38,10%	34,77%	6,80%
16	39,68%	41,27%	29,10%	34,92%	39,68%	34,92%	40,21%	37,11%	4,03%
17	83,60%	79,37%	64,02%	76,19%	75,66%	83,07%	81,48%	77,63%	6,26%
18	50,26%	54,50%	30,69%	56,08%	57,67%	54,50%	53,97%	51,10%	8,59%
Média	**45,18%**	**46,44%**	**35,04%**	**45,34%**	**51,77%**	**50,67%**	**48,03%**		
Desvio Padrão	**13,89%**	**13,54%**	**12,75%**	**12,93%**	**10,40%**	**13,58%**	**13,42%**		

Nesta mesma visão de análise, observou-se que a medida LCSS apresentou o melhor desvio padrão com 10,40% (Figura 11).

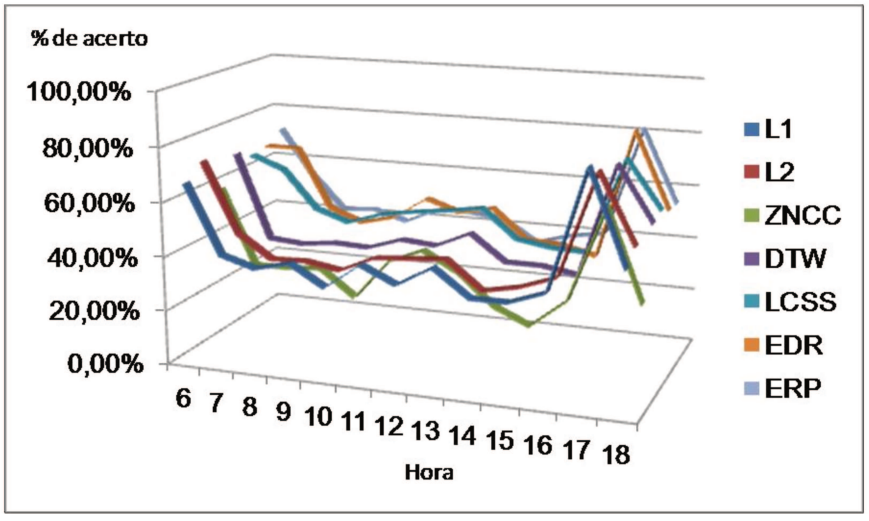

Figura 11: % Acerto 1-NN - Espécie: *Aspidosperma tomentosum*

Espécie: *Miconia rubiginosa*

Na visão por espécie *Miconia rubiginosa,* identificou-se na Tabela 9 que a medida de distância L1 apresentou o melhor percentual de acerto com 34,87% e que as 18 horas apresentou o melhor percentual de acerto com 40,63%.

Tabela 9: Análise por Espécie – *Miconia rubiginosa*

Hora	L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
6	28,89%	27,78%	6,67%	25,56%	23,33%	27,78%	27,78%	23,97%	7,27%
7	33,33%	33,33%	16,67%	25,56%	24,44%	20,00%	26,67%	25,71%	5,77%
8	18,89%	18,89%	14,44%	15,56%	8,89%	16,67%	17,78%	15,87%	3,24%
9	36,67%	40,00%	18,89%	38,89%	20,00%	25,56%	38,89%	31,27%	8,74%
10	36,67%	35,56%	15,56%	41,11%	33,33%	32,22%	37,78%	33,17%	7,68%
11	36,67%	31,11%	25,56%	31,11%	23,33%	32,22%	30,00%	30,00%	4,07%
12	24,44%	28,89%	31,11%	26,67%	24,44%	23,33%	30,00%	26,98%	2,83%
13	28,89%	28,89%	36,67%	28,89%	21,11%	24,44%	30,00%	28,41%	4,48%
14	47,78%	46,67%	31,11%	30,00%	38,89%	36,67%	46,67%	39,68%	6,97%
15	44,44%	45,56%	25,56%	45,56%	31,11%	33,33%	47,78%	39,05%	8,17%
16	31,11%	28,89%	27,78%	34,44%	32,22%	30,00%	28,89%	30,48%	2,13%
17	44,44%	43,33%	24,44%	48,89%	34,44%	35,56%	42,22%	39,05%	7,57%
18	41,11%	43,33%	27,78%	44,44%	43,33%	42,22%	42,22%	40,63%	5,34%
Média	**34,87%**	**34,79%**	**23,25%**	**33,59%**	**27,61%**	**29,23%**	**34,36%**		
Desvio Padrão	**8,09%**	**8,10%**	**7,99%**	**9,30%**	**8,75%**	**6,88%**	**8,56%**		

Nesta mesma visão de análise, observou-se que a medida EDR apresentou o melhor desvio padrão com 6,88 % (Figura 12).

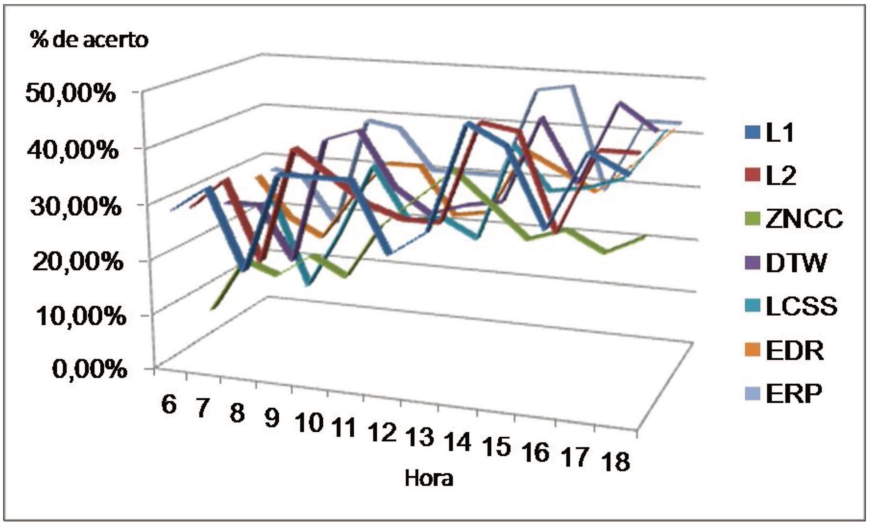

Figura 12: % Acerto 1-NN - Espécie: *Miconia rubiginosa*

Espécie: *Pouteria ramiflora*

Na visão por espécie *Pouteria ramiflora,* identificou-se na Tabela 10 que a medida de distância ZNCC apresentou o melhor percentual de acerto com 9,14% e que as 07 horas apresentou o melhor percentual de acerto com 15,58%.

Tabela 10: Análise por Espécie – *Pouteria ramiflora*

Hora	L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
6	4,24%	7,27%	12,73%	10,30%	2,42%	3,64%	9,09%	7,10%	3,54%
7	12,73%	12,73%	36,97%	16,36%	7,88%	9,70%	12,73%	15,58%	9,07%
8	9,09%	9,70%	15,15%	16,36%	6,06%	7,88%	10,30%	10,65%	3,49%
9	4,85%	3,03%	6,67%	3,03%	2,42%	1,82%	4,24%	3,72%	1,53%
10	4,24%	3,64%	3,64%	2,42%	2,42%	3,03%	4,24%	3,38%	0,71%
11	1,21%	1,82%	4,24%	6,67%	0,00%	0,00%	0,61%	2,08%	2,31%
12	6,06%	7,27%	12,73%	9,70%	4,85%	5,45%	8,48%	7,79%	2,56%
13	3,64%	4,24%	5,45%	3,64%	4,24%	5,45%	4,24%	4,42%	0,70%
14	1,82%	1,82%	1,82%	4,24%	1,21%	0,61%	1,82%	1,90%	1,05%
15	2,42%	1,82%	1,82%	4,85%	1,21%	1,82%	2,42%	2,34%	1,10%
16	4,24%	4,24%	6,67%	5,45%	3,03%	3,64%	4,24%	4,50%	1,12%
17	6,06%	6,06%	4,24%	9,09%	2,42%	6,06%	9,09%	6,15%	2,23%
18	14,55%	16,36%	6,67%	13,94%	9,70%	11,52%	15,76%	12,64%	3,26%
Média	**5,78%**	**6,15%**	**9,14%**	**8,16%**	**3,68%**	**4,66%**	**6,71%**		
Desvio Padrão	**3,89%**	**4,32%**	**9,00%**	**4,73%**	**2,68%**	**3,34%**	**4,38%**		

Nesta mesma visão de análise, observou-se que a medida LCSS apresentou o melhor desvio padrão com 2,68% (Figura 13).

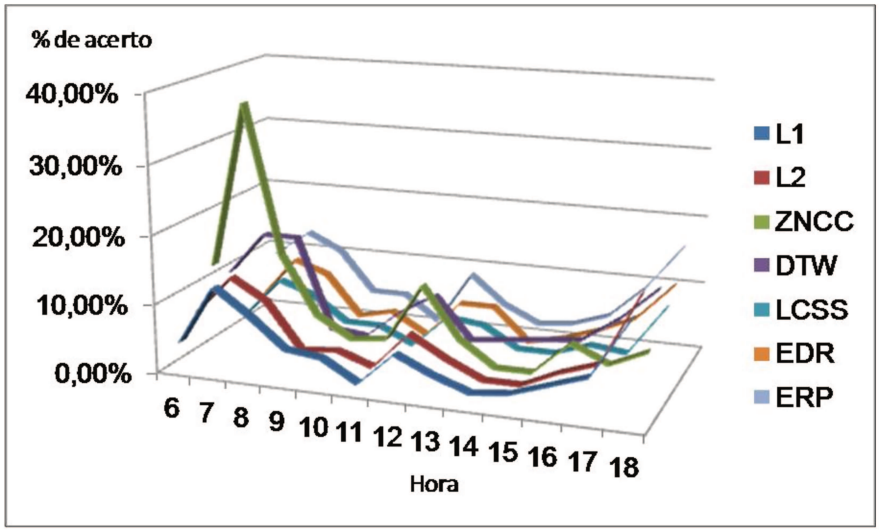

Figura 13: % Acerto 1-NN - Espécie: *Pouteria ramiflora*

4.3.4 Análise considerando canal de cor por espécie e por hora.

Canal de Cor *Red* por Espécie

Na Tabela 11 observou-se que para a espécie *Aspidosperma tomentosum*, a medida ZNCC apresentou o melhor percentual de acerto com 28,08%, para as 17 horas com 83,22% e a medida DTW apresentou o melhor desvio padrão com 15,09%.

Para a espécie *Miconia rubiginosa* a medida ERP apresentou o melhor percentual de acerto com 53,59%, para as 17 horas com 69,05% e a medida ZNCC apresentou o melhor desvio padrão com 6,86%.

Para a espécie *Pouteria ramiflora* a medida ZNCC apresentou o melhor percentual de acerto com 16,78%, para as 18 horas com 35,06%, além disso, a medida LCSS apresentou o melhor desvio padrão com 6,97%.

Tabela 11: Análise por Canal de Cor *Red* e Espécie

Cor: ***Red***

Espécie	Hora	Medidas de Distância L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
Aspidosperma tomentosum	6	30,16%	41,27%	25,40%	36,51%	31,75%	25,40%	41,27%	33,11%	6,26%
	7	4,76%	3,17%	0,00%	4,76%	44,44%	50,79%	11,11%	17,01%	19,67%
	8	11,11%	12,70%	23,81%	19,05%	15,87%	19,05%	15,87%	16,78%	3,97%
	9	12,70%	9,52%	26,98%	12,70%	20,63%	17,46%	17,46%	16,78%	5,42%
	10	17,46%	19,05%	15,87%	23,81%	23,81%	19,05%	22,22%	20,18%	2,90%
	11	19,05%	17,46%	42,86%	23,81%	26,98%	28,57%	19,05%	25,40%	8,14%
	12	12,70%	14,29%	36,51%	25,40%	25,40%	23,81%	19,05%	22,45%	7,47%
	13	19,05%	17,46%	30,16%	23,81%	25,40%	26,98%	14,29%	22,45%	5,26%
	14	14,29%	11,11%	23,81%	9,52%	14,29%	12,70%	11,11%	13,83%	4,39%
	15	12,70%	11,11%	19,05%	19,05%	14,29%	11,11%	11,11%	14,06%	3,33%
	16	17,46%	15,87%	31,75%	11,11%	7,94%	6,35%	12,70%	14,74%	7,86%
	17	87,30%	85,71%	87,30%	65,08%	79,37%	90,48%	87,30%	83,22%	8,04%
	18	4,76%	4,76%	1,59%	7,94%	26,98%	20,63%	3,17%	9,98%	9,09%
	Média	**20,27%**	**20,27%**	**28,08%**	**21,73%**	**27,47%**	**27,11%**	**21,98%**		
	Desvio Padrão	**20,35%**	**20,88%**	**20,72%**	**15,09%**	**17,39%**	**21,03%**	**20,69%**		

Cor: ***Red***

Espécie	Hora	Medidas de Distância L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
Miconia rubiginosa	6	63,33%	56,67%	6,67%	63,33%	60,00%	66,67%	56,67%	53,33%	19,35%
	7	66,67%	73,33%	30,00%	53,33%	46,67%	36,67%	63,33%	52,86%	14,85%
	8	13,33%	23,33%	16,67%	20,00%	10,00%	23,33%	23,33%	18,57%	4,99%
	9	56,67%	60,00%	13,33%	66,67%	23,33%	40,00%	56,67%	45,24%	18,76%
	10	56,67%	46,67%	6,67%	66,67%	50,00%	43,33%	53,33%	46,19%	17,59%
	11	63,33%	43,33%	10,00%	46,67%	26,67%	40,00%	40,00%	38,57%	15,42%
	12	33,33%	40,00%	20,00%	23,33%	30,00%	26,67%	43,33%	30,95%	7,91%
	13	13,33%	13,33%	20,00%	23,33%	10,00%	10,00%	20,00%	15,71%	4,95%
	14	63,33%	66,67%	16,67%	33,33%	56,67%	60,00%	66,67%	51,90%	17,89%
	15	63,33%	63,33%	13,33%	70,00%	46,67%	56,67%	73,33%	55,24%	18,93%
	16	43,33%	46,67%	3,33%	50,00%	50,00%	43,33%	46,67%	40,48%	15,37%
	17	83,33%	83,33%	13,33%	90,00%	60,00%	70,00%	83,33%	69,05%	24,61%
	18	73,33%	73,33%	20,00%	76,67%	80,00%	76,67%	70,00%	67,14%	19,47%
	Média	**53,33%**	**53,08%**	**14,62%**	**52,56%**	**42,31%**	**45,64%**	**53,59%**		
	Desvio Padrão	**20,75%**	**19,41%**	**6,86%**	**21,45%**	**20,10%**	**18,83%**	**18,00%**		

Cor: ***Red***

Espécie	Hora	Medidas de Distância L1	L2	ZNCC	DTW	LCSS	EDR	ERP	Média	Desvio Padrão
Pouteria ramiflora	6	5,45%	9,09%	30,91%	16,36%	3,64%	5,45%	9,09%	11,43%	8,84%
	7	16,36%	18,18%	61,82%	18,18%	14,55%	10,91%	14,55%	22,08%	16,39%
	8	14,55%	20,00%	32,73%	36,36%	7,27%	12,73%	21,82%	20,78%	9,81%
	9	10,91%	7,27%	10,91%	3,64%	5,45%	3,64%	9,09%	7,27%	2,92%
	10	9,09%	9,09%	9,09%	3,64%	1,82%	3,64%	9,09%	6,49%	3,05%
	11	3,64%	5,45%	10,91%	14,55%	0,00%	0,00%	1,82%	5,19%	5,18%
	12	7,27%	7,27%	16,36%	9,09%	3,64%	3,64%	7,27%	7,79%	3,97%
	13	3,64%	3,64%	9,09%	3,64%	3,64%	7,27%	3,64%	4,94%	2,11%
	14	1,82%	1,82%	1,82%	5,45%	1,82%	0,00%	1,82%	2,08%	1,51%
	15	7,27%	5,45%	5,45%	9,09%	3,64%	1,82%	7,27%	5,71%	2,26%
	16	7,27%	7,27%	9,09%	7,27%	3,64%	5,45%	7,27%	6,75%	1,60%
	17	12,73%	10,91%	7,27%	14,55%	3,64%	7,27%	12,73%	9,87%	3,62%
	18	41,82%	47,27%	12,73%	38,18%	27,27%	32,73%	45,45%	35,06%	11,20%
	Média	**10,91%**	**11,75%**	**16,78%**	**13,85%**	**6,15%**	**7,27%**	**11,61%**		
	Desvio Padrão	**9,86%**	**11,40%**	**15,66%**	**11,08%**	**6,97%**	**8,19%**	**11,06%**		

Canal de Cor *Green* por Espécie

Na tabela 12 observou-se que para a espécie *Aspidosperma tomentosum*, a medida LCSS apresentou o melhor percentual de acerto com 79,00%, para as 6 horas com 88,66 % e a medida LCSS apresentou o melhor desvio padrão com 6,32%.

Para a espécie *Miconia rubiginosa* a medida ZNCC apresentou o melhor percentual de acerto com 33,85%, para as 18 horas com 25,24%, a medida EDR apresentou o melhor desvio padrão com 3,46%.

Para a espécie *Pouteria ramiflora* a medida DTW apresentou o melhor percentual de acerto com 3,92%, para as 7 horas com 10,39% e a medida LCSS apresentou o melhor desvio padrão com 1,36%.

Tabela 12: Análise por Canal de Cor *Green e* Espécie

Cor:	*Green*								
		Medidas de Distância							
Espécie	**Hora**	**L1**	**L2**	**ZNCC**	**DTW**	**LCSS**	**EDR**	**ERP**	**Média**	**Desvio Padrão**
Aspidosperma tomentosum	6	88,89%	90,48%	84,13%	88,89%	88,89%	90,48%	88,89%	88,66%	1,98%
	7	73,02%	76,19%	77,78%	73,02%	79,37%	77,78%	76,19%	76,19%	2,24%
	8	71,43%	74,60%	42,86%	60,32%	88,89%	79,37%	71,43%	69,84%	13,63%
	9	69,84%	66,67%	28,57%	61,90%	66,67%	61,90%	65,08%	60,09%	13,13%
	10	55,56%	52,38%	11,11%	60,32%	74,60%	71,43%	53,97%	54,20%	19,32%
	11	61,90%	61,90%	19,05%	60,32%	73,02%	77,78%	63,49%	59,64%	17,65%
	12	57,14%	63,49%	34,92%	57,14%	73,02%	69,84%	63,49%	59,86%	11,56%
	13	68,25%	69,84%	33,33%	66,67%	84,13%	74,60%	66,67%	66,21%	14,59%
	14	60,32%	60,32%	17,46%	65,08%	76,19%	74,60%	60,32%	59,18%	18,15%
	15	74,60%	73,02%	19,05%	61,90%	76,19%	76,19%	79,37%	65,76%	19,75%
	16	73,02%	69,84%	26,98%	63,49%	79,37%	74,60%	69,84%	65,31%	16,29%
	17	82,54%	74,60%	66,67%	85,71%	82,54%	80,95%	73,02%	78,00%	6,26%
	18	82,54%	80,95%	52,38%	80,95%	84,13%	84,13%	80,95%	78,00%	10,54%
	Média	**70,70%**	**70,33%**	**39,56%**	**68,13%**	**79,00%**	**76,43%**	**70,21%**		
	Desvio Padrão	**9,78%**	**9,41%**	**22,93%**	**10,16%**	**6,32%**	**6,67%**	**9,13%**		

Cor:	*Green*								
		Medidas de Distância							
Espécie	**Hora**	**L1**	**L2**	**ZNCC**	**DTW**	**LCSS**	**EDR**	**ERP**	**Média**	**Desvio Padrão**
Miconia rubiginosa	6	6,67%	10,00%	3,33%	3,33%	0,00%	3,33%	10,00%	5,24%	3,50%
	7	0,00%	0,00%	3,33%	0,00%	3,33%	3,33%	0,00%	1,43%	1,65%
	8	3,33%	3,33%	23,33%	3,33%	3,33%	3,33%	3,33%	6,19%	7,00%
	9	3,33%	3,33%	26,67%	3,33%	3,33%	3,33%	3,33%	6,67%	8,16%
	10	6,67%	6,67%	30,00%	6,67%	3,33%	6,67%	6,67%	9,52%	8,44%
	11	0,00%	0,00%	53,33%	0,00%	0,00%	0,00%	0,00%	7,62%	18,66%
	12	0,00%	0,00%	56,67%	3,33%	0,00%	0,00%	0,00%	8,57%	19,67%
	13	0,00%	3,33%	43,33%	0,00%	0,00%	3,33%	3,33%	7,62%	14,66%
	14	6,67%	3,33%	43,33%	0,00%	0,00%	0,00%	3,33%	8,10%	14,57%
	15	6,67%	10,00%	40,00%	6,67%	3,33%	0,00%	10,00%	10,95%	12,31%
	16	3,33%	3,33%	53,33%	3,33%	6,67%	3,33%	3,33%	10,95%	17,34%
	17	6,67%	3,33%	23,33%	13,33%	3,33%	3,33%	3,33%	8,10%	7,10%
	18	20,00%	26,67%	40,00%	36,67%	13,33%	13,33%	26,67%	25,24%	9,74%
	Média	**4,87%**	**5,64%**	**33,85%**	**6,15%**	**3,08%**	**3,33%**	**5,64%**		
	Desvio Padrão	**5,17%**	**6,84%**	**16,84%**	**9,50%**	**3,57%**	**3,46%**	**6,84%**		

Cor:	*Green*								
		Medidas de Distância							
Espécie	**Hora**	**L1**	**L2**	**ZNCC**	**DTW**	**LCSS**	**EDR**	**ERP**	**Média**	**Desvio Padrão**
Pouteria ramiflora	6	3,64%	5,45%	1,82%	3,64%	1,82%	3,64%	9,09%	4,16%	2,32%
	7	10,91%	10,91%	12,73%	12,73%	3,64%	7,27%	14,55%	10,39%	3,47%
	8	1,82%	0,00%	1,82%	1,82%	0,00%	0,00%	0,00%	0,78%	0,90%
	9	3,64%	1,82%	3,64%	1,82%	0,00%	0,00%	1,82%	1,82%	1,37%
	10	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
	11	0,00%	0,00%	1,82%	3,64%	0,00%	0,00%	0,00%	0,78%	1,32%
	12	5,45%	9,09%	5,45%	5,45%	3,64%	3,64%	9,09%	5,97%	2,11%
	13	1,82%	1,82%	0,00%	3,64%	0,00%	0,00%	1,82%	1,30%	1,27%
	14	1,82%	1,82%	0,00%	1,82%	0,00%	1,82%	1,82%	1,30%	0,82%
	15	0,00%	0,00%	0,00%	3,64%	0,00%	1,82%	0,00%	0,78%	1,32%
	16	1,82%	1,82%	5,45%	3,64%	1,82%	1,82%	0,00%	2,34%	1,60%
	17	1,82%	3,64%	1,82%	5,45%	0,00%	5,45%	5,45%	3,38%	2,05%
	18	0,00%	0,00%	3,64%	3,64%	1,82%	1,82%	0,00%	1,56%	1,51%
	Média	**2,52%**	**2,80%**	**2,94%**	**3,92%**	**0,98%**	**2,10%**	**3,36%**		
	Desvio Padrão	**2,90%**	**3,47%**	**3,39%**	**2,93%**	**1,36%**	**2,24%**	**4,56%**		

Canal de Cor *Blue* por Espécie

Na tabela 13 observou-se que para a espécie *Aspidosperma tomentosum*, a medida ERP apresentou o melhor percentual de acerto com 51,89%, para as 6 horas com 80,27 %, a medida LCSS apresentou o melhor desvio padrão com 13,78%.

Para a espécie *Miconia rubiginosa* a medida L1 apresentou o melhor percentual de acerto com 46,41%, para as 13 horas com 61,90 % e a medida ZNCC apresentou o melhor desvio padrão com 11,66%.

Para a espécie *Pouteria ramiflora* a medida ZNCC apresentou o melhor percentual de acerto com 7,69%, para as 7 horas com 14,29 %, a medida L2 apresentou o melhor desvio padrão com 3,26%.

Tabela 13: Análise por Canal de Cor *Blue* e Espécie

Cor: ***Blue***

		Medidas de Distância								
Espécie	**Hora**	**L1**	**L2**	**ZNCC**	**DTW**	**LCSS**	**EDR**	**ERP**	**Média**	**Desvio Padrão**
Aspidosperma tomentosum	6	80,95%	84,13%	66,67%	82,54%	77,78%	85,71%	84,13%	80,27%	6,05%
	7	46,03%	58,73%	14,29%	34,92%	60,32%	73,02%	74,60%	51,70%	20,06%
	8	30,16%	25,40%	25,40%	30,16%	36,51%	34,92%	34,92%	31,07%	4,23%
	9	39,68%	38,10%	39,68%	39,68%	41,27%	39,68%	42,86%	40,14%	1,40%
	10	26,98%	36,51%	39,68%	28,57%	42,86%	38,10%	38,10%	35,83%	5,42%
	11	47,62%	46,03%	49,21%	41,27%	47,62%	47,62%	49,21%	46,94%	2,53%
	12	41,27%	50,79%	55,56%	39,68%	53,97%	47,62%	49,21%	48,30%	5,55%
	13	46,03%	44,44%	42,86%	49,21%	49,21%	47,62%	46,03%	46,49%	2,20%
	14	30,16%	30,16%	31,75%	36,51%	36,51%	28,57%	31,75%	32,20%	2,90%
	15	17,46%	25,40%	17,46%	30,16%	30,16%	26,98%	23,81%	24,49%	4,94%
	16	28,57%	38,10%	28,57%	30,16%	31,75%	23,81%	38,10%	31,29%	4,85%
	17	80,95%	77,78%	38,10%	77,78%	65,08%	77,78%	84,13%	71,66%	14,75%
	18	63,49%	77,78%	38,10%	79,37%	61,90%	58,73%	77,78%	65,31%	13,67%
	Média	**44,57%**	**48,72%**	**37,48%**	**46,15%**	**48,84%**	**48,47%**	**51,89%**		
	Desvio Padrão	**19,15%**	**19,37%**	**14,01%**	**19,29%**	**13,78%**	**19,20%**	**20,11%**		

Cor: ***Blue***

		Medidas de Distância								
Espécie	**Hora**	**L1**	**L2**	**ZNCC**	**DTW**	**LCSS**	**EDR**	**ERP**	**Média**	**Desvio Padrão**
Miconia rubiginosa	6	16,67%	16,67%	10,00%	10,00%	10,00%	13,33%	16,67%	13,33%	3,09%
	7	33,33%	26,67%	16,67%	23,33%	23,33%	20,00%	16,67%	22,86%	5,47%
	8	40,00%	30,00%	3,33%	23,33%	13,33%	23,33%	26,67%	22,86%	10,90%
	9	50,00%	56,67%	16,67%	46,67%	33,33%	33,33%	56,67%	41,90%	13,67%
	10	46,67%	53,33%	10,00%	50,00%	46,67%	46,67%	53,33%	43,81%	14,08%
	11	46,67%	50,00%	13,33%	46,67%	43,33%	56,67%	50,00%	43,81%	13,02%
	12	40,00%	46,67%	16,67%	53,33%	43,33%	43,33%	46,67%	41,43%	10,82%
	13	73,33%	70,00%	46,67%	63,33%	53,33%	60,00%	66,67%	61,90%	8,70%
	14	73,33%	70,00%	33,33%	56,67%	60,00%	50,00%	70,00%	59,05%	13,06%
	15	63,33%	63,33%	23,33%	60,00%	43,33%	43,33%	60,00%	50,95%	13,88%
	16	46,67%	36,67%	26,67%	50,00%	40,00%	43,33%	36,67%	40,00%	7,13%
	17	43,33%	43,33%	36,67%	43,33%	40,00%	33,33%	40,00%	40,00%	3,56%
	18	30,00%	30,00%	23,33%	20,00%	36,67%	36,67%	30,00%	29,52%	5,75%
	Média	**46,41%**	**45,64%**	**21,28%**	**42,05%**	**37,44%**	**38,72%**	**43,85%**		
	Desvio Padrão	**15,61%**	**16,35%**	**11,66%**	**16,41%**	**13,91%**	**13,37%**	**17,09%**		

Cor: ***Blue***

		Medidas de Distância								
Espécie	**Hora**	**L1**	**L2**	**ZNCC**	**DTW**	**LCSS**	**EDR**	**ERP**	**Média**	**Desvio Padrão**
Pouteria ramiflora	6	3,64%	7,27%	5,45%	10,91%	1,82%	1,82%	9,09%	5,71%	3,29%
	7	10,91%	9,09%	36,36%	18,18%	5,45%	10,91%	9,09%	14,29%	9,69%
	8	10,91%	9,09%	10,91%	10,91%	10,91%	10,91%	9,09%	10,39%	0,82%
	9	0,00%	0,00%	5,45%	3,64%	1,82%	1,82%	1,82%	2,08%	1,80%
	10	3,64%	1,82%	1,82%	3,64%	5,45%	5,45%	3,64%	3,64%	1,37%
	11	0,00%	0,00%	0,00%	1,82%	0,00%	0,00%	0,00%	0,26%	0,64%
	12	5,45%	5,45%	16,36%	14,55%	7,27%	9,09%	9,09%	9,61%	3,97%
	13	5,45%	7,27%	7,27%	3,64%	9,09%	9,09%	7,27%	7,01%	1,80%
	14	1,82%	1,82%	3,64%	5,45%	1,82%	0,00%	1,82%	2,34%	1,60%
	15	0,00%	0,00%	0,00%	1,82%	0,00%	1,82%	0,00%	0,52%	0,82%
	16	3,64%	3,64%	5,45%	5,45%	3,64%	3,64%	5,45%	4,42%	0,90%
	17	3,64%	3,64%	3,64%	7,27%	3,64%	5,45%	9,09%	5,19%	2,05%
	18	1,82%	1,82%	3,64%	0,00%	0,00%	0,00%	1,82%	1,30%	1,27%
	Média	**3,92%**	**3,92%**	**7,69%**	**6,71%**	**3,92%**	**4,62%**	**5,17%**		
	Desvio Padrão	**3,48%**	**3,26%**	**9,30%**	**5,21%**	**3,41%**	**4,01%**	**3,63%**		

4.3.5 Análise de correlação entre as medidas de distância.

A acurácia obtidas das análises anteriores e de outras possíveis podem ainda proporcionar o benefício da análise das correlações entre as medidas, o que permite averiguar que certas medidas de distância podem ser substituídas por outras sem prejuízos nos resultados obtidos.

No trabalho proposto por (SANTOS et al., 2012), os autores desenvolveram um descritor para avaliar a correlação entre classificadores onde utiliza uma matriz de relacionamento para calcular a pontuação de correlação de pares de classificadores. Neste trabalho, especificamente, o descritor foi aplicado nas medidas de distância.

Na análise de correlação dessas medidas de distância investigadas, constatou-se conforme sugerido nos gráficos anteriores, e apresentado na matriz de correlação na Tabela 14, que o percentual de correlação entre as medidas de distância L1, L2, DTW, LCSS, EDR e ERP são altos tanto para o 1-NN, quanto para os demais KNNs. Para o 1-NN por exemplo, o percentual de correlação da medida de distância L1 para a medida L2 é de 86%, da medida L1 para a medida DTW é de 76%, para a medida ERP é de 85%, para a medida EDR é de 74%, para a medida LCSS é de 71% e da medida L2 para a medida ERP é de 94%. Por outro lado, o percentual de correlação das demais medidas para a ZNCC é baixo, especificamente para a medida L1 é de 31%. Esta análise indica que as medidas L1, L2, e ERP possuem resultados semelhantes, possibilitando a escolha de uma delas para outros trabalhos relacionados. Estes resultados podem guiar a escolha por técnicas de fusão de classificadores que possam ser utilizadas para combinar classificadores com boa eficácia que sejam menos correlacionados.

Tabela 14 – Matriz de Correlação

			KNN 1						
			L1	L2	ZNCC	DTW	LCSS	EDR	ERP
			1	2	3	4	5	6	7
KNN 1	L1	1	1	0,86	0,31	0,76	0,71	0,74	0,85
	L2	2		1	0,33	0,73	0,71	0,74	0,94
	ZNCC	3			1	0,28	0,28	0,28	0,33
	DTW	4				1	0,66	0,67	0,72
	LCSS	5					1	0,83	0,71
	EDR	6						1	0,75
	ERP	7							1

4.4 Considerações finais

Neste capítulo foram apresentados os procedimentos e os resultados para a avaliação da acurácia das medidas de distância envolvendo 4 visões de análise, sendo análise geral por horário, por canal de cor, por espécie e por canal de cor, espécie e hora. Além dessas visões de análise, foi apresentada a análise de correlação entre as medidas de distância.

Na visão geral por horário, identificou-se que a medida de distância ERP apresentou o melhor percentual médio de acerto e que as 17 horas apresentou o melhor percentual médio de acerto, independente da medida de similaridade avaliada.

Na visão por canal de cor *Red,* identificou-se que a medida de distância DTW apresentou o melhor percentual médio de acerto e que para as séries referentes às 17 horas apresentou o melhor percentual médio de acerto, independente da medida de similaridade.

Na visão por canal de cor *Green,* identificou-se que a medida de distância LCSS apresentou o melhor percentual médio de acerto e que para as séries referentes às 6 horas apresentou o melhor percentual médio de acerto, independente da medida de similaridade.

Na visão por canal de cor *Blue,* identificou-se que a medida de distância ERP apresentou o melhor percentual médio de acerto e que para as séries referentes às 17 horas apresentou o melhor percentual médio de acerto, independente da medida de similaridade.

Na visão por espécie *Aspidosperma tomentosum,* identificou-se que a medida de distância LCSS apresentou o melhor percentual médio de acerto e que para as séries referentes às 17 horas apresentou o melhor percentual médio de acerto, independente da medida de similaridade.

Na visão por espécie *Miconia rubiginosa,* identificou-se que a medida de distância L1 apresentou o melhor percentual médio de acerto e que para as séries referentes às 18 horas apresentou o melhor percentual médio de acerto, independente da medida de similaridade.

Na visão por espécie *Pouteria ramiflora,* identificou-se que a medida de distância ZNCC apresentou o melhor percentual médio de acerto e que para as séries referentes às 7 horas apresentou o melhor percentual médio de acerto, independente da medida de similaridade.

Na análise de correlação, observou-se que as medidas de distância L1, L2 e ERP possuem comportamentos parecidos.

Os resultados apresentados em todas as visões de análise indicam que os percentuais mais altos de acerto acontecem nas primeiras horas do dia e no final da tarde.

Os indicadores mostram também que a medida de distância ERP apresentou os melhores resultados e a medida ZNCC apresentou os piores resultados, conforme observado na Tabela 15.

Tabela 15 – Visão geral dos resultados

Cor	Espécie	Hora	L1	L2	DTW	ZNCC	LCSS	EDR	ERP
Geral	Geral	**17 (+)** **10 (-)**				-			+
Red	Geral	**17 (+)** **13 (-)**			+	-			
	Aspidosperma Tomentosum	**17 (+)** **18 (-)**	-	-		+			
	Miconia Rubiginosa	**17 (+)** **13 (-)**				-			+
	Pouteria Ramiflora	**18 (+)** **14 (-)**				+	-		
Green	Geral	**6 (+)** **10 (-)**				-	+		
	Aspidosperma Tomentosum	**6 (+)** **10 (-)**				-	+		
	Miconia Rubiginosa	**18 (+)** **7 (-)**				+	-		
	Pouteria Ramiflora	**7 (+)** **10 (-)**			+		-		
Blue	Geral	**17 (+)** **15 (-)**				-			+
	Aspidosperma Tomentosum	**6 (+)** **15 (-)**				-			+
	Miconia Rubiginosa	**13 (+)** **6 (-)**	+			-			
	Pouteria Ramiflora	**7 (+)** **11 (-)**	-	-		+	-		
Geral	*Aspidosperma Tomentosum*	**17 (+)** **15 (-)**				-	+		
	Miconia Rubiginosa	**18 (+)** **8 (-)**	+			-			
	Pouteria Ramiflora	**7 (+)** **14 (-)**				+	-		

5 Conclusões

Esse trabalho de mestrado se concentrou na investigação de medidas de distância para a classificação de séries temporais em arquivos contendo características de imagens de vegetação, nas quais pretendeu identificar a partir dos dados analisados informações relevantes para a área de fenologia.

A medida de distância ERP apresentou os melhores percentuais de acerto, por outro lado, a medida ZNCC apresentou os piores percentuais de acerto.

Na análise de correlação, observou-se que as medidas de distância L1, L2 e ERP apresentaram resultados parecidos.

No trabalho preliminar em que os rótulos desconhecidos são considerados no cálculo de similaridade pelas medidas de distância, o percentual de acurácia se apresenta mais baixo. Esse percentual baixo de acurácia pode ter acontecido dado a grande quantidade de indivíduos de espécies desconhecidas, devido a diversidade de vegetação do cerrado.

A Figura 14 mostra o gráfico representado por hora e percentual de acerto para as medidas L1, L2 e ZNCC.

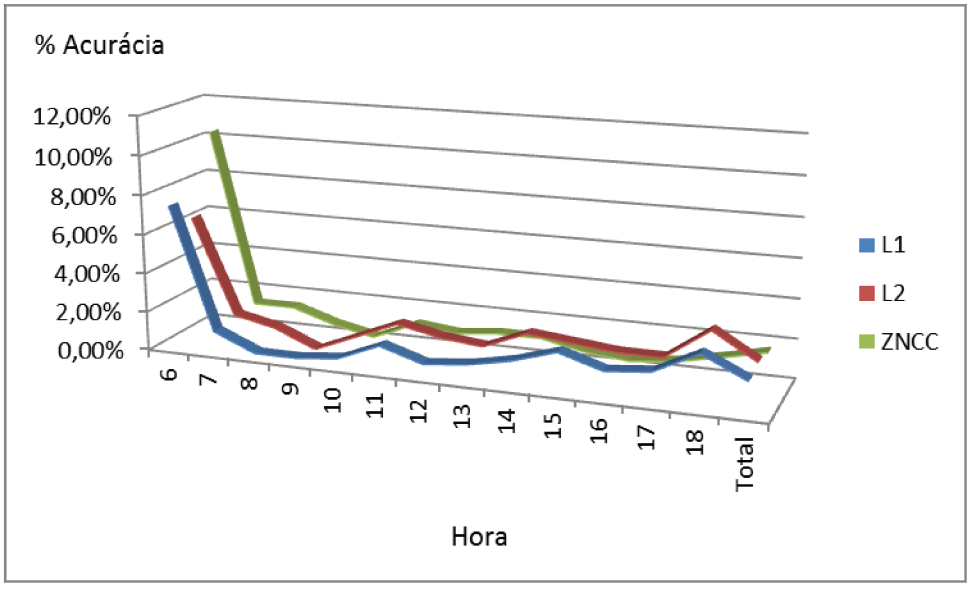

Figura 14 – % Acerto com rótulos desconhecidos

5.1 Contribuições do Trabalho

A contribuição do trabalho está na realização da análise de eficácia de medidas de distância em dados da área de fenologia de plantas. Ressaltando que não há trabalhos semelhantes publicados.

5.2 Trabalhos Futuros

Como proposta de possíveis trabalhos futuros, poderíamos citar:

- A inclusão dos resultados obtidos com as medidas de similaridade no banco de dados do *e-phenology*;
- Analisar outras espécies de planta;
- Analisar os indivíduos de espécies desconhecidas;
- Analisar as medidas de similaridade sob a ótica de grupos funcionais Decídua (perdem as folhas no outono, renovando na primavera), Semidecidua (não perdem totalmente as folhas durante o ano) e sempre verde. Esses grupos funcionais foram propostos por (Almeida et al., 2012).

5.3 Considerações finais

Este trabalho teve resultados esperados, pois analisou a eficácia das medidas de distância para a classificação de séries temporais associadas ao comportamento fenológico de plantas.

A pesquisa em fenologia pode se beneficiar com os resultados obtidos neste trabalho no sentido de buscar identificar se existem semelhanças entre as espécies de plantas avaliadas e quais os fatores que favorecem essas semelhanças em determinados horários do dia em que estas espécies foram registradas pela câmera digital.

Referências Bibliográficas

ALBERTI, L. F.; MORELLATO, L. P. C. Influência da abertura de trilhas antrópicas e clareiras naturais na fenologia reprodutiva de Gymnanthes concolor (Spreng .) Müll . Arg . (Euphorbiaceae). Revista Brasileira de Botânica, v. 31, n. 1, p. 53-59, 2008.

ALMEIDA, J., DOS SANTOS, J.A.; ALBERTON, B.; TORRES, R. da S.; MORELLATO, L.P.C., Remote phenology: Applying machine learning to detect phenological patterns in a Cerrado Savanna. IEEE International Conference on eScience, p.1-8, 2012.

ALMEIDA, J.; DOS SANTOS, J.; ALBERTON, B.; TORRES, R. da S.; MORELLATO, L. P. C. Applying machine learning based on multiscale classifiers to detect remote phenology patterns in Cerrado savanna trees. Ecological Informatics, jul. 2013.

ANDRIENKO, G.; ANDRIENKO, N.; MLADENOV, M.; MOCK, M.; POELITZ, C. Extracting Events from Spatial Time Series. IEEE 14th International Conference on Information Visualisation, p. 48-53, 2010.

BATISTA, G. E. A. P. A.; WANG, X.; KEOGH, E. J. A Complexity-Invariant Distance Measure for Time Series. 11th SIAM International Conference on Data Mining, p. 699-710, 2011.

BERNDT, D. and CLIFFORD, J. Using Dynamic Time Warping to Find Patterns in Time Series. AAAI Workshop on Knowledge Discovery in Databases, p. 359-370, 1994.

BROCKWELL, P. J.; DAVIS, R. A. Introduction to Time Series and Forecasting, 2nd Edition, Springer Texts in Statistics, 1997

CASTRO, E. R.; GALETTI, M.; MORELLATO, L. P. C. Reproductive phenology of Euterpe edulis (Arecaceae) along a gradient in the Atlantic rainforest of Brazil. Australian Journal of Botany, v. 55, n. 7, p. 725-735, 2007.

CHAMBERS, L. E.; ALTWEGG, R.; BARBRAUD, et al. Phenological changes in the southern hemisphere. PloS one, v. 8, n. 10, jan 2013.

CHAOVALITWONGSE, W. A.; FAN, Y.J.; SACHDEO, R. C.. On the Time Series K - Nearest Neighbor Classification of Abnormal Brain Activity. IEEE Transactions on Systems, v. 37, n. 6, p. 1005-1016, 2007.

CHEN, L.; NG, R. On The Marriage of Lp-norms and Edit Distance. p. 792-803, VLDB, 2004.

CHEN, L.; OZSU, M. T.; ORIA, V. Robust and Fast Similarity Search for Moving Object Trajectories. SIGMOD Conference , p. 491-502, 2005.

COUTO Jr, A. F.; CARVALHO Jr; O. A. de; et al.. Tratamento de Ruídos e Caracterização de Fisionomias do Cerrado Uitlizando Séries Temporais do Sensor Modis. Revista Árvore, v.35, n.3, Edição Especial, p. 699-705, 2011.

DING, H.; TRAJCEVSKI, G.; SCHEUERMANN, P.; WANG, X.; KEOGH, E. Querying and Mining of Time Series Data : Experimental Comparison of Representations and Distance Measures. VLDB, v. 1, n. 212, 2008.

ESLING, P.; AGON, C.; RECHERCHE, I. DE. Time-Series Data Mining. ACM Computing Surveys, v. 45, n. 1, 2012.

FALOUTSOS, C.; RANGANATHAN, M.; MANOLOPOULOS, Y. Fast Subsequence Matching in Time-Series Databases. ACM SIGMOD International Conference on Management of Data, v. 23, n. 2, p. 419-429, 1994.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. Artificial Intelligence Magazine, v. 17, n. 3, 1996.

FERREIRA, M. R. P.; SANTOS, L. F. D.; TRAINA, A. J. M.; DIAS, I.; CHBEIR, R.; TRAINA Jr., C. Algebraic Properties to Optimize kNN Queries. Journal of Information and Data Management, v. 2, n. 3, 2011.

GUYON , D.; GUILLOT, M.; VITASSE, Y.; CARDOT, H.; HAGOLLE; O.; DELZON, S. AND WIGNERON, J.-P. Remote Sensing of Environment Monitoring elevation variations in leaf phenology of deciduous broadleaf forests from Spot/Vegetation time-series. Elsevier, Remote Sensing of Environment, 2010.

HAN, J.; KAMBER, M. Data Mining – Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

HUDSON, I. L. ; KIM, S. W. AND KEATLEY, M. R. Climatic influences on the flowering phenology of four Eucalypts : a GAMLSS approach. 18th World IMACS/MODSIM Congress, July, p. 2611-2617, 2009.

KASHYAP, S.; KARRAS, P. Scalable k NN Search on Vertically Stored Time Series ∗ Categories and Subject Descriptors. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, v. 1, p. 1334-1342, 2011.

KEOGH, E. Exact Indexing of Dynamic Time Warping. VLDB Conference. 2002.

MACKUTE-VARONECKIENE, A.; ZILINSKAS, A.; VARONECKAS, A. Multidimensional Scaling: Multi-Objective Optimization Approach. ACM CompSysTech, p. 1-6, 2009.

MARIOTE, L. E.; MEDEIROS, C. B.; TORRES, R. da S.; BUENO, L. M. TIDES - a new descriptor for time series oscillation behavior. Geoinformatica, Ed. Springer, p. 75-109, 2011.

MARTIN J. and CROWLEY J. L. Experimental comparison of correlation techniques, Int. Conf. on Intelligent Autonomous Systems, 1995.

MCCLOY, K. R. Development and Evaluation of Phenological Change Indices Derived from Time Series of Image Data. Remote Sensing, p. 2442-2473, 2010.

MEZER, A.; YOVEL, Y.; PASTERNACK, O.; GORNE, T. and ASSAF, Y. Cluster analysis of resting-state fmri time series. Neuroimage, v. 45, n. 4, p. 1117–1125, 2009.

MONARD, M. C. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACM SIGKDD, v. 6, n. 1, 2004.

MORELLATO, E. P. C.; TORRES, R. da S. http://www.recod.ic.unicamp.br/ephenology/index.php?id=2. Acessado em 30/06/2012.

MORELLATO, L. P. C.; CAMARGO, M. G. G. and GRESSLER, E. Chapter 6 A Review of Plant Phenology in South and Central America, Springer. 2013.

MUEEN, A.; NATH, S. and LIU, J. Fast Approximate Correlation for Massive Time-series Data. SIGMOD 2010, v. 1, p. 171-182, 2010.

OLIVEIRA, P. C. de. Séries Temporais : Analisar o Passado , Predizer o Futuro. Analysis, p. 3-6, 2007.

OTRANTO, E. Identifying financial time series with similar dynamic conditional correlation. Computational Statistics & Data Analysis 54, 115, 2010.

PINTO, A. M.; MORELLATO, L. P. C.; BARBOSA, A. P. Willd (Fabaceae) em duas áreas de floresta na Amazônia Central. ACTA AMAZONICA, v. 38, n. 4, p. 643-650, 2000.

RATHCKE, B.;LACEY, E. P. Phenological patterns of terrestrial plants, Annual Review of Ecology and Systematics, n.16, p. 179–214, 1985.

RICHARDSON, A. D.; JENKINS, J. P.; BRASWELL, B. H.; HOLLINGER, D. Y.; OLLINGER, S. V. and SMITH, M-L. . Use of digital webcam images to track spring green-up in a deciduous broadleaf forest. Oecologia, v. 152, n. 2, p. 323-34, 2007.

ROMANI, L. A. S.; AVILA, A. M. H.; ZULLO Jr., J.; TRAINA Jr., C. and TRAINA, A. J. M. Mining Relevant and Extreme Patterns on Climate Time Series with CLIPSMiner. America, v. 1, n. 2, p. 245-260, 2010.

SANTOS, J. A. dos; FARIA, F. A.; TORRES, R. da S.; ROCHA, A.; GOSSELIN, P-H.; PHILIPP-FOLIGUET, S. and FALCÃO, A. Descriptor Correlation Analysis for Remote Sensing Image Multi-Scale Classification, 21st International Conference on Pattern Recognition (ICPR), p. 3078-3081, 2012.

SCHWARTZ, M.D.Phenology: An Integrative Environmental Science. Kluwer Academic Publishers, 2003.

SCHNEIDER, T.; O'GORMAN, P. A.; LEVINE, X. Water vapor and the dynamics of climate changes. Reviews of Geophysics. v. 48, n. 3, p. 1-23, 2009.

SHANMUGANATHAN, S.; SALLIS, P. and NARAYANAN, A. Modelling the seasonal climate effects on grapevine yield at different spatial and unconventional temporal scales. International Congress on Environmental Modeling and Software, 2010.

TAN, P; STEINBACH, M. and KUMAR, V. Introduction to Data Mining. Addison Wesley, 2005.

UENO, K.; LEE, D. Anytime Classification Using the Nearest Neighbor Algorithm with Applications to Stream Mining. Sixth IEEE International Conference on Data Mining, 2006.

UOTILA, P; LYNCH, A. H.; CASSANO, J. J. and CULLATHER, R. I. Changes in Antarctic net precipitation in the 21st century based on Intergovernmental Panel on Climate Change (IPCC) model scenarios. Journal of Geophysical Research, v. 12, n. 10, 2007.

VLACHOS, M.; KOLLIOS, G.; GUNOPULOS, D. Discovering similar multidimensional trajectories. 18th International Conference on Data Engineering, p. 673-684, 2002.

WEI, L.-L. Mining Regression-Classes in Fuzzy Point Data Sets. Sixth International Conference on Fuzzy Systems and Knowledge Discovery, p. 562-566, 2009.

YI, B. K. Yi AND FALOUTSOS, C. Fast Time Sequence Indexing for Arbitrary Lp Norms. VLDB. 2000.

ZHANG, M.; FAN, J.; ZHU, X.; LI, G. and ZHANG, Y.. Monitoring winter-wheat phenology in North China using time-series. Agriculture, v. 7472, p. 1-6, 2009.

ZHAO, J.; ZHANG, Y.; TAN, Z.; SONG, Q.; LIANG, N.; YU, L. and ZHAO, J. Ecological Informatics Using digital cameras for comparative phenological monitoring in an evergreen broad-leaved forest and a seasonal rain forest. Ecological Informatics, v. 10, p. 65–72, 2012.