

Daniel Hideki Shibuya

Avaliação de técnicas de preenchimentos de dados faltantes em séries temporais de imagens do satélite VENμS para classificação do uso da terra

Trabalho de conclusão de curso apresentado como requisito parcial para obtenção do título de **Bacharel em Ciências Agrícolas** à Faculdade de engenharia agrícola da Universidade Estadual de Campinas.

Orientadora: Gleyce K.D.A Figueiredo

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Shibuya, Daniel Hideki, 1996-
Sh61a Avaliação de técnicas de preenchimento de dados faltantes em séries
temporais de imagens do satélite VENUS para classificação do uso da terra /
Daniel Hideki Shibuya. – Campinas, SP : [s.n.], 2021.

Orientador: Gleyce Kelly Dantas Araujo Figueiredo.
Coorientador: Gisela Maiara Silva Pereira.
Trabalho de Conclusão de Curso (graduação) – Universidade Estadual de
Campinas, Faculdade de Engenharia Agrícola.

1. Interpolação. 2. Geoprocessamento. 3. Satélites artificiais em agricultura. 4.
Mapeamento digital. 5. Floresta aleatória. I. Figueiredo, Gleyce Kelly Dantas
Araujo, 1984-. II. Pereira, Gisela Maiara Silva. III. Universidade Estadual de
Campinas. Faculdade de Engenharia Agrícola. IV. Título.

Informações adicionais complementares

Título em outro idioma: Evaluation of missing data filling techniques in time series of VENUS
satellite images for land use classification

Palavras-chave em inglês:

Interpolation
Geoprocessing
Artificial satellites in agriculture
Digital mapping
Random Forest

Titulação: Bacharel

Banca examinadora:

Julianne de Castro Oliveira
Ana Claudia dos Santos Luciano

Data de entrega do trabalho definitivo: 12-01-2021

Avaliação de técnicas de preenchimentos de dados faltantes em séries temporais de imagens do satélite VENμS para classificação do uso da terra

Daniel Hideki Shibuya

BANCA EXAMINADORA

Gleyce Kelly Dantas Araújo Figueiredo

Julianne de Castro Oliveira

Ana Claudia dos Santos Luciano

Agradecimentos

Os dados do VEN μ S foram adquiridos pela Centre National d'Etudes Spatiales (CNES), pelas agências da França e Israel. A área de estudo faz parte do Joint Experiment of Crop Assessment and Monitoring (JECAM) e também agradecemos ao Eder Araújo da Silva (<http://www.floragroapoio.com.br>) pelo suporte técnico. Agradeço também a Faculdade de Engenharia Agrícola (FEAGRI) que forneceu a infraestrutura e o conhecimento ao longo de toda a minha graduação e que possibilitaram esse projeto de Trabalho de Conclusão de Curso.

Agradeço também os membros da banca examinadora pelas correções e sugestões desse trabalho. O pesquisador Guerric le Maire, que contribuiu fornecendo os materiais necessários para realização do estudo. Finalizo agradecendo minha orientadora Gleyce K.D.A Figueiredo e co-orientadora Gisela M.S Perreira por todo o ensinamento que me passaram durante o período do projeto.

Resumo

A aplicação do sensoriamento remoto na agricultura para a elaboração de mapas de classificação do uso da terra representa uma alternativa para monitoramento de culturas agrícolas. O monitoramento agrícola a partir de dados remotos, costuma ser feito através de imagens de satélite ao longo do tempo. No Brasil, dificuldades causadas pela alta presença de nuvens e consequentemente uma descontinuidade na série temporal causam problemas de interpretação da mesma. Afim de evitar problemas causados pela descontinuidade de uma série temporal, existem métodos que permitem o preenchimento dos dados faltantes. Este trabalho apresenta um estudo de 5 métodos diferentes de preenchimento de dados faltantes, que são: Interpolação, Média, Mediana e *MissForest* utilizando séries temporais do satélite experimental *Vegetation and Environment monitoring on New Micro Satellite* (VENμS). A área de estudo é caracterizada pela alta dinâmica de culturas ao longo do tempo, e fica localizada entre os municípios de Avaré e Botucatu. O índice de vegetação *Normalized Difference Vegetation Index* – NDVI foi calculado com o objetivo de auxiliar a interpretação da série. Além disso, este trabalho também verificou o potencial de cada método de preenchimento para classificação orientada a objeto utilizando o algoritmo *Random Forest*. Os métodos de preenchimento, foram avaliados através de análise visual dos perfis temporais e estatística do preenchimento e matriz de confusão e suas métricas gerados pelo classificador *Random Forest*. O método de preenchimento que apresentou melhor resultado foi o de Interpolação linear ($k=1$) apresentando boa adaptação a culturas com alta dinâmica ao longo do tempo, como a cana-de-açúcar. Esse mesmo banco de dados com Interpolação $k=1$ atingiu acurácia de 0,81 e permitiu a elaboração de um mapa a atividade de classificação, destacando a importância de preencher os dados faltantes presentes na série temporal.

Palavras-Chaves: Interpolação, NDVI, Classificação Orientada a Objeto, *Random Forest*.

Abstract

The application of remote sensing in agriculture for the elaboration of land use classification maps represents an alternative for monitoring agricultural crops. Agricultural monitoring from remote data is usually done through satellite images over time. In Brazil, difficulties caused by the high presence of clouds and consequently a discontinuity in the time series cause problems in its interpretation. In order to avoid problems caused by the discontinuity of a time series, there are methods that allow filling in the missing data. This work presents a study of five different methods of filling in missing data, which are Interpolation, Average, Median and *MissForest* using time series from the experimental satellite Vegetation and Environment monitoring on New Micro Satellite (VEN μ S). A study area is characterized by the high dynamics of cultures over time, and is located between the municipalities of Avare and Botucatu. The Normalized Difference Vegetation Index - NDVI vegetation index was calculated in order to assist in the interpretation of the series. In addition, this work also verified the potential of each filling method for object-oriented classification using the Random Forest algorithm. The filling methods were evaluated through visual analysis of the temporal profiles and statistics of the filling and confusion matrix and its metrics generated by the *Random Forest* classifier. The filling method that presented the best result was that of linear interpolation ($k = 1$) showing good adaptation to cultures with high dynamics over time, such as sugarcane. This same database with Interpolation $k = 1$ reached an accuracy of 0.81 and allowed the elaboration of a map to the classification activity, highlighting the importance of filling in the missing data present in the time series.

Keywords: Interpolation, NDVI, Object Based Image Analysis, Random Forest.

Sumário

1 Introdução	1
1.1 justificativa	2
1.2 Objetivos	2
2 Revisão Bibliográfica	3
2.1 Monitoramento Agrícola	3
2.2 Satélite VENμS	3
2.3 Dados faltantes em séries temporais	5
2.4 Preenchimento de dados faltantes	5
2.5 Classificação do uso da terra	6
3 Material e Métodos	7
3.1 Área de Estudo e dados de campo	7
3.2 Dados Espectrais	8
3.3 Pré-processamento das imagens	8
3.4 Segmentação	11
3.5 Índice de vegetação	11
3.6 Preenchimento dos dados faltantes	11
3.7 Validação Cruzada	12
3.8 Classificação do uso da terra	14
4 Resultados e Discussão	16
4.1 Preenchimento dos Dados Faltantes	16
4.2 Classificação do uso da terra	18
5 Conclusão	24
6 Referências Bibliográficas	26
Apêndice:	30

1 Introdução

Uma série temporal dados é definida como um conjunto de observações de um dado ordenado no tempo. Existem diferentes tipos de aplicações para as séries temporais, podendo-se mencionar seu uso na agricultura para detectar mudanças ao longo do tempo, ou para monitoramento agrícola contínuo (ARÉVALO et al., 2020; DIAO, 2019). O que torna importante a análise de séries temporais são as observações às quais as séries temporais podem ser associadas. Segundo BAYER e SOUZA (2010), o objetivo de compreender a dependência entre tempo e dados observados é essencial para identificar comportamentos e fazer previsões.

O uso de imagens de satélites multitemporal para criação de séries temporais, mostraram grande crescimento, sendo amplamente utilizado como auxílio no monitoramento agrícola (SANTOS et al., 2019; WERNER et al., 2020). Os avanços das capacidades de imageamento dos sensores, processamento de dados e também a facilidade de acesso gratuito às imagens de satélite são fatores que contribuíram para o seu crescimento (ESQUERDO, 2016).

O sensor *Moderate Resolution Imaging Spectroradiometer* – MODIS é um dos mais utilizados para aplicações de séries temporais na agricultura, sendo usado em diversos estudos. BORGES et.al (2014) utilizou dados do MODIS para mapear o uso e a cobertura vegetal no oeste da Bahia. WERNER et al. (2020) utilizou o sensor MODIS para construir um modelo para identificar e separar culturas com comportamento espectral a de uma série temporal. O MODIS apresenta uma boa resolução temporal de 2 dias, no entanto ele tem limitações principalmente devido à baixa resolução espacial que varia de 250 a 1000m (LATORRE, 2003).

Estudos como o de LUCIANO et.al (2019) utilizaram os satélites Landsat 7 e 8 afim de evitar a limitação da resolução temporal do MODIS para mapear a cultura de cana-de-açúcar no estado de São Paulo. BELGIU (2017) fez uso do satélite Sentinel-2 e destaca a importância de uma boa resolução temporal para a atividade de classificação por meio do *Random Forest*. O satélite *Vegetation and Environment monitoring on a New Micro Satellite* (VEN μ S), é um satélite experimental desenvolvido pelas agências espaciais francesa e israelense e um dos objetivos dessa missão é oferecer imagens de alta resolução temporal com 2 dias de revisita além de um conjunto de 12 bandas espectrais cujo comprimento de onda varia de 0,420 μ m até 0,910 μ m e resolução espacial de 5,3m (HERSCOVITZ et al., 2018). Apesar dos grandes avanços de sensores acoplados em satélites como o VEN μ S, regiões tropicais ainda sofrem com ausência de dados em função de condições meteorológicas, causando lacunas na série temporal.

Essas lacunas ou dados faltantes nas séries temporais apresentam um desafio ao pré-processamento. A cobertura de nuvens, nevoeiro e sombras de nuvens são as causas mais comuns dessa falta de dados. Em regiões tropicais, como o Brasil, devido às características do clima, essa ausência de dados é maior (HOLLOWAY et al., 2019).

A ausência de dados provocadas pela presença de nuvens dificulta a identificação de padrões dos alvos na superfície, sendo esse um problema cuja complexidade não se encontra apenas na extração de dados, mas também nos métodos utilizados para preencher os dados faltantes (CARVALHO 2017). Como os métodos de análise de séries temporais foram e são desenvolvidos, em grande parte, para conjunto de dados completos, mesmo uma pequena quantidade de dados ausentes pode causar um problema nas estimativas, justificando a importância de analisar os dados faltantes (HAREL, 2007).

1.1 justificativa

Problemas de descontinuidade de séries temporais de imagens de satélite ainda são um grande desafio. Devido à dificuldade de interpretar a série temporal quando há a presença de dados faltantes, a busca por métodos que possibilite amenizar essas lacunas é a melhor alternativa para usuários dessas séries temporais.

1.2 Objetivos

Este trabalho tem como objetivo avaliar diferentes métodos de preenchimento de dados faltantes na série temporal de imagens de NDVI do satélite VEN μ S. Para alcançar esse objetivo, os seguintes objetivos específicos foram traçados, que são:

- Testar os métodos de preenchimento de dados faltantes de Média, Mediana, *MissForest* e interpolação ($k=1$ e $k=2$);
- Avaliar o uso das séries temporais de imagens preenchidas a partir de cada método para a classificação orientada a objeto do uso e cobertura da terra utilizando o algoritmo *Random Forest*.

2 Revisão Bibliográfica

2.1 Monitoramento Agrícola

De acordo com a EMBRAPA (2014), o monitoramento agrícola pode ser definido como um estudo dinâmico ao longo do tempo em uma determinada área, tendo como objetivo acompanhar o uso da terra por meio de imagens de satélite. No caso do Brasil, existem plataformas que disponibilizam imagens de forma gratuita, facilitando acesso a esses dados. Para os produtores rurais, o monitoramento do uso da terra representa uma alternativa de aumentar a produção (CHINO et.al 2010).

O monitoramento agrícola através do sensoriamento remoto possui diversas aplicações, devido a quantidade de informações que podem ser extraídas das imagens de satélites, pode-se citar exemplos como o relevo da área, umidade do solo e também índices de vegetação. Um dos índices mais utilizados mundialmente é o *Normalized Difference Vegetation Index* (NDVI), que permite identificar a presença de vegetação verde e sua distribuição espacial (LUCIANO et.al 2018).

Gerar mapas de produtividade é um dos principais exemplos de aplicação do monitoramento agrícola, e tem como objetivo identificar dentro da área de estudo regiões que apresentam baixa ou alta produtividade dentro da lavoura. Essa prática permite que o produtor tenha maior eficiência no manejo de insumos como adubo e fertilizantes, e assim diminuir os custos de produção (POVH et.al 2008).

Mapas de classificação obtidos através do sensoriamento remoto, podem ser aplicados de diversas maneiras de forma a ajudar a interpretação dos dados fornecidos pelas imagens de satélite. LUCIANO et.al (2018), em seu trabalho utilizou séries temporais aliadas ao método de *machine learning* permitindo mapear áreas produtoras de cana-de-açúcar em larga escala. Além disso, as informações fornecidas pelas imagens de satélite permitem também mapear áreas com mais de uma cultura, incluindo áreas de integração lavoura-pecuária (MANABE et.al 2018).

2.2 Satélite VEN μ S

O satélite *Vegetation and Environment monitoring on a New Micro Satellite* (VEN μ S) foi desenvolvido em conjunto pelas agências espaciais francesa e israelense, com o objetivo do monitoramento da vegetação da Terra que leva a bordo o sensor VEN μ S *Superspectral Camera* (VSSC). Foi lançado em agosto de 2017, é um microssatélite síncrono-solar quase polar a uma inclinação de 98 °, orbitando a uma altitude de 720 km. O sensor cobre uma área de 27 km com um ângulo de visão constante (HERSCOVITZ et al., 2018).

O satélite possui resolução temporal de 2 dias, podendo coletar imagens de todo o globo (HERSCOVITZ et al., 2018). De acordo com descrição do catálogo do satélite VEN μ S, o objetivo científico do satélite é o fornecimento de dados para estudos científicos que lidam com o monitoramento, análise e modelagem do funcionamento da superfície terrestre sob a influência de fatores ambientais e atividades humanas. Dessa forma, seu uso e aplicação para monitoramento de atividades agrícolas está de acordo com o seu objetivo.

Dentre as principais informações disponíveis, há um conjunto de 12 bandas espectrais estreitas cujo comprimento de onda varia de 0,420 μ m até 0,910 μ m e resolução espacial de 5,3m (Tabela 1). As imagens do sensor VSSC possuem alta resolução espacial de 5,3 metros, porém corrigidas para 10 metros, após correção atmosférica.

Bandas	Comprimentos de onda central (μm)	Comprimento da banda (nm)	Objetivo
B1	0.420	40	Correção atmosférica
B2	0.443	40	Correção de nuvens
B3	0.490	40	Correção atmosférica
B4	0.555	40	Índices de vegetação
B5	0.620	40	Índices de vegetação
B6	0.620	40	Correção de elevação
B7	0.667	30	Índices de vegetação
B8	0.702	24	Índices de vegetação
B9	0.742	16	Índices de vegetação
B10	0.782	16	Índices de vegetação
B11	0.865	40	Índices de vegetação
B12	0.910	20	Correção de copos d'agua

Tabela 1: Especificações do satélite VEN μ S, adaptado de HERSCOVITZ et al., 2018.

Apesar de ser um satélite experimental, existem alguns estudos explorando a capacidade do satélite VEN μ S. HAGOLLE et.al (2015), desenvolveu um método de correção atmosférica para imagens obtidas pelo VEN μ S. LIAO et.al (2020) combinou dados do satélite VEN μ S com dados de radar para a classificação multitemporal de culturas agrícolas e destaca que o uso do satélite auxiliou nos resultados obtidos para a classificação.

2.3 Dados faltantes em séries temporais

O uso de séries temporais representa um método utilizado para a análise de dados em um determinado período de tempo. ANTUNES (2015), descreve a série temporal como uma sequência de dados quantitativos relativos a um momento específico e estudados utilizando sua distribuição com o tempo. Variáveis como umidade do solo e uso do solo são exemplos de como aplicar a série temporal para atividades agrícolas (POGGIO et.al, 2012).

No entanto o uso de séries temporais é suscetível a problemas como a presença de dados faltantes, o que prejudica toda a análise do conjunto de dados (WEISS et.al, 2014). Dessa forma, a presença de dado faltante numa série temporal voltada para a agricultura, também pode ser considerada como um fator limitante para a análise e assim provocar possíveis erros de interpretação dos dados.

O monitoramento de áreas agrícolas é muito comum fazer uso de imagens de sensoriamento remoto para o acompanhamento da lavoura, permitindo que o agricultor tenha uma avaliação de sua área como um todo, através de índices de vegetação, por exemplo (SHIRATSUCHI et.al, 2014).

No sensoriamento remoto, a causa mais comum para a presença de dados faltantes é a presença de nuvens, principalmente quando o alvo do estudo são culturas de verão. Ao se trabalhar com uma série temporal de dados orbitais, a presença de dados relativos a nuvem prejudica a análise de comportamento da superfície, dessa forma as regiões da imagem que apresentam nuvens passam a ser tratados como dados faltante. O Brasil, é um país que em boa parte do seu território apresenta características de clima tropical, ou seja, existe alto índice de chuvas durante o ano inteiro, o que aumenta probabilidade de ocorrência de dados faltantes durante uma série temporal (MARTINELLI ,2010)

2.4 Preenchimento de dados faltantes

Independentemente da origem do dado faltante, em séries temporais a principal alternativa é o preenchimento dessa série utilizando métodos estatísticos que possam torná-lo o mais próximo possível de um valor real. Existem diversos estudos que tem como objetivo avaliar esses métodos de preenchimento (WEISS et.al, 2014, POGGIO et.al, 2012), buscam avaliar técnicas de preenchimento de dados (imputação) por meio dos pixels presente nas imagens de seus respectivos bancos de dados com dados faltantes.

O preenchimento de dados faltantes pode ser classificado como simples ou múltipla. Na imputação simples, o dado é preenchido uma única vez de acordo com o método utilizado. Já na imputação múltipla, são feitas diversas imputações para um mesmo dado faltante, possibilitando a inclusão da incerteza de cada método presente no resultado, sendo esta a principal vantagem da imputação múltipla sob a imputação simples (NUNES et.al, 2009).

Os métodos mais comuns para a realização da imputação de dados são pelos modelos estatísticos de média, moda, mediana e interpolação, sendo a última mais recomendada para séries temporais pois leva em consideração a variação da série ao longo do tempo (PRATAMA et.al 2016). Métodos de imputação simples e múltipla, podem ser encontrados em plataformas como *Rstudio* e *Python*. Com objetivo de utilizar os dados preenchidos para um problema de classificação, SANTOS et.al (2019) realizou o preenchimento dos dados faltantes em uma série temporal do Landsat 7 ETM⁺ e Landsat 8 OLI para classificar o uso da terra em uma região no estado São Paulo e concluiu que a imputação simples foi suficiente para essa aplicação.

O software livre *RStudio* (R CORE TEAM, 2020), apresenta diversos tipos diferentes de pacotes gratuitos que permitem o preenchimento de dados faltantes. O pacote *imputeTS* (*Time Series Missing Value Imputation*) (MORITZ, 2017), é um pacote específico para preenchimento de dados faltante de séries temporais. Nesse pacote há diferentes métodos de preenchimento de dados como: Média, Mediana e Interpolação linear com o parâmetro $k=1$ e $k=2$. Sendo o parâmetro k o número de dados que serão utilizados para fazer a interpolação linear (número de vizinhos).

É possível também pacotes que utilizam métodos iterativos como o *MissForest* (STEKHOFEN. 2013), que é utilizado para preenchimento de dados lineares e não lineares, em dados categórico ou contínuo, porém não é específico para séries temporais. O método de imputação utilizado neste pacote é baseado no conceito de árvores de decisão, uma metodologia que é muito utilizada por apresentar vantagens como a fácil interpretação e apresentar um raciocínio de classificação semelhante ao cérebro humano (JAMES, 2013).

2.5 Classificação do uso da terra

A classificação do uso da terra é uma atividade que permite a identificação de diferentes elementos presentes num determinado ambiente. Esse tipo de classificação apresenta diversos usos para o monitoramento agrícola e ambiental (CHINO et.al 2010). Diversos estudos utilizam a classificação do uso da terra como metodologia de trabalho. GASPARINI et.al (2013) utilizou a classificação para a identificação de áreas de preservação em propriedades agrícolas. LUCIANO et.al (2018), também utilizaram a classificação para monitorar a cana-de-açúcar no estado de São Paulo.

Existem diversas maneiras de se realizar a classificação, em seus respectivos trabalhos LUCIANO et.al (2018) e SANTOS et.al (2019) destacam a eficácia do algoritmo *Random Forest*. Trata-se de um algoritmo classificador/regressor não paramétrico, que utiliza múltiplas árvores de decisões para a obtenção de resultados, apresentando bom comportamento para bancos de dados com muitas informações (BREIMAN, 2001).

Assim como muitas metodologias, o *Random Forest* não atua na presença de dados faltantes, sendo uma dificuldade para o seu uso em países como o Brasil e também justificando a necessidade de um método eficiente de preenchimento de dados (SANTOS et.al 2019).

3 Material e Métodos

3.1 Área de Estudo e dados de campo

A área de estudo deste trabalho está localizada no centro do estado de São Paulo, entre os municípios de Avaré e Botucatu (Figura 1) com um total de 1700 km². A região de estudo foi selecionada em função da abrangência do satélite VENμS na área, permitindo a presença de diversas culturas diferentes. A área apresenta alta diversidade de culturas (anual e perene) com plantações de eucaliptos, pinheiros, cítricos e vegetação natural. O estudo compreende o período entre novembro de 2017 a outubro de 2019.

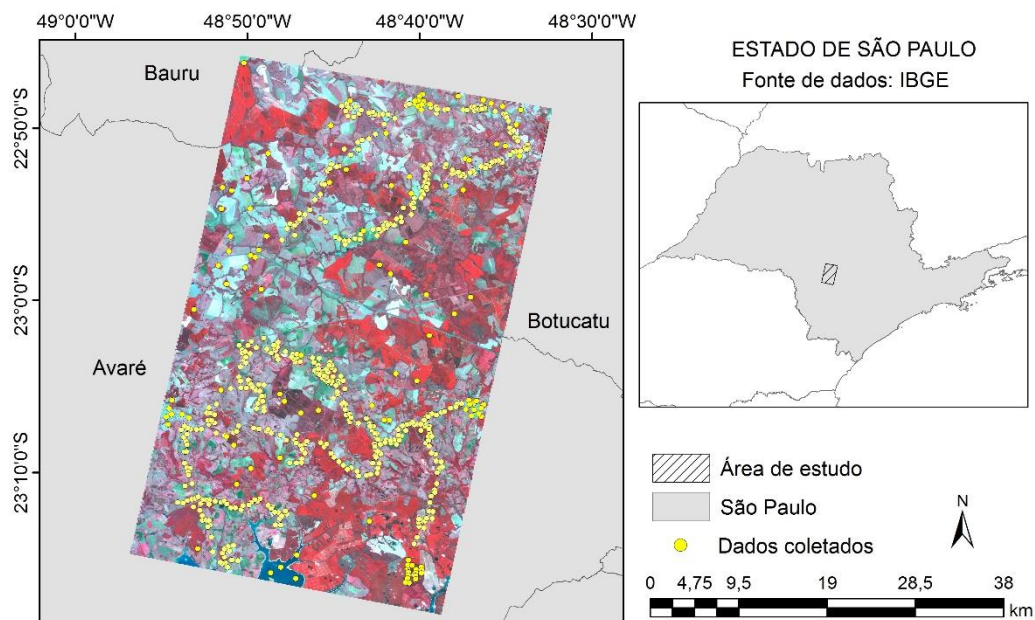


Figura 1: Mapa referente a área de estudo e respectivos pontos de coleta de dados. Imagem obtida através do Satélite VENμS no dia 13 de setembro de 2018, bandas utilizadas: RGB 11-5-4.

O clima na região é classificado como subtropical úmido (ROLIM; APARECIDO, 2015). O solo é classificado como latossolo, e o relevo é considerado suave (encostas na maioria das vezes <5%), com altitude entre 500 e 700 m. A temperatura média é aproximadamente 19°C, enquanto a precipitação média anual é de 1400 mm.

A coleta de dados de campo foi realizada por meio do acompanhamento do uso da terra. Essa coleta foi feita *in loco* para a verificação da cultura que está instalada no momento e fotografar a área de interesse.

A cada coleta de dados de campo são coletados 463 pontos ao longo da estrada, correspondentes a 463 polígonos definidos manualmente utilizando imagem do VENμS (área de interesse de 44,82 km²). Os polígonos englobam áreas de alta dinâmica com a rotação de cultura, lagos e áreas urbanas. Essa coleta foi realizada a cada dois meses para acompanhar a rotação das culturas. Os polígonos são rotulados de acordo com a cultura que está presente na área.

Além disso, 109 pontos foram coletados manualmente com a ajuda da ferramenta *Google Earth*. A classificação foi feita visualmente e teve como objetivo identificar áreas urbanas e corpos d'água presentes na área de estudo. Dessa forma, o número total de amostras foi de 572.

3.2 Dados Espectrais

Neste trabalho os dados espectrais são extraídos do satélite VENμS. A resolução espacial utilizada foi de 10 metros, com resolução temporal de 2 dias. As imagens são tratadas na França, país de origem do satélite, através do algoritmo de código aberto *MAJA* responsável pela detecção de nuvens e também pela correção atmosférica das imagens. No banco de dados espectrais do satélite VENμS para o presente estudo há 172 imagens inseridas no período de estudo (novembro de 2017 a outubro de 2019).

3.3 Pré-processamento das imagens

Apesar da disponibilização de uma máscara de nuvem, foi necessário fazer um refinamento nesse produto para obtenção de um dado de qualidade. Assim, para todas as imagens da série temporal com nuvens foram feitas máscaras utilizando a banda 2 do satélite VENμS e assim indicar a presença de nuvens, sombra e névoa, utilizando o *software* ArcMap (Figura 2). Além disso, foi feito um refinamento manual da máscara de forma a encontrar o maior número de regiões a serem retiradas. Esses dados retirados são considerados como dado faltante e através destas informações foi quantificado a qualidade de pixels em cada uma das imagens (Figura 3 e Figura 4).

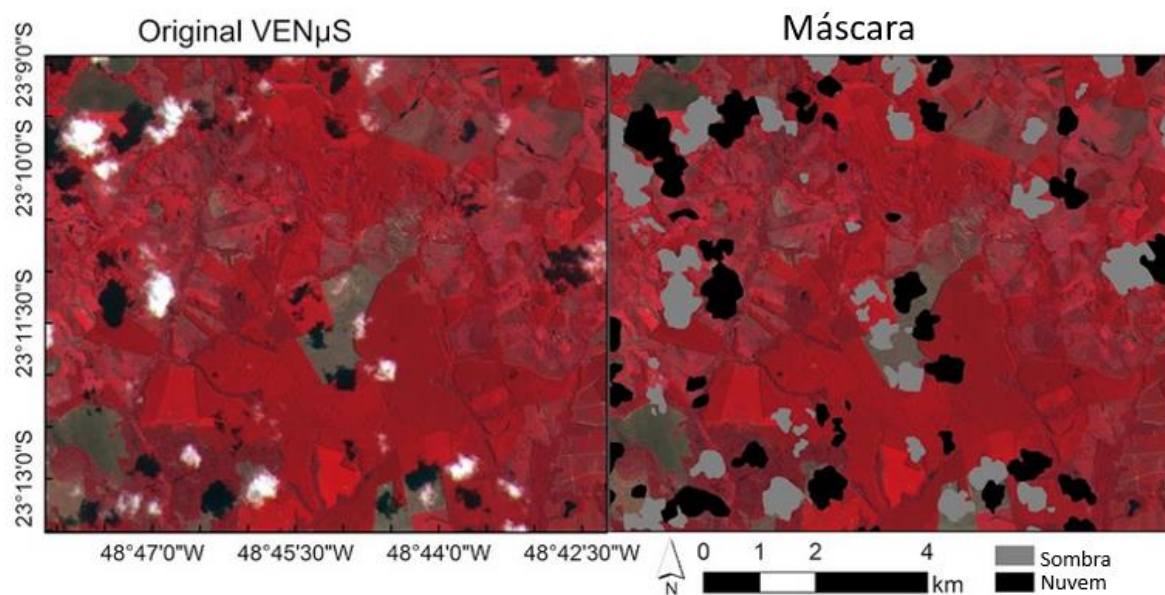


Figura 2: Imagem original sem a retirada de nuvens, e imagem com o tratamento da máscara para regiões com nuvem e sombra de nuvem. Composição RGB 11-5-4.

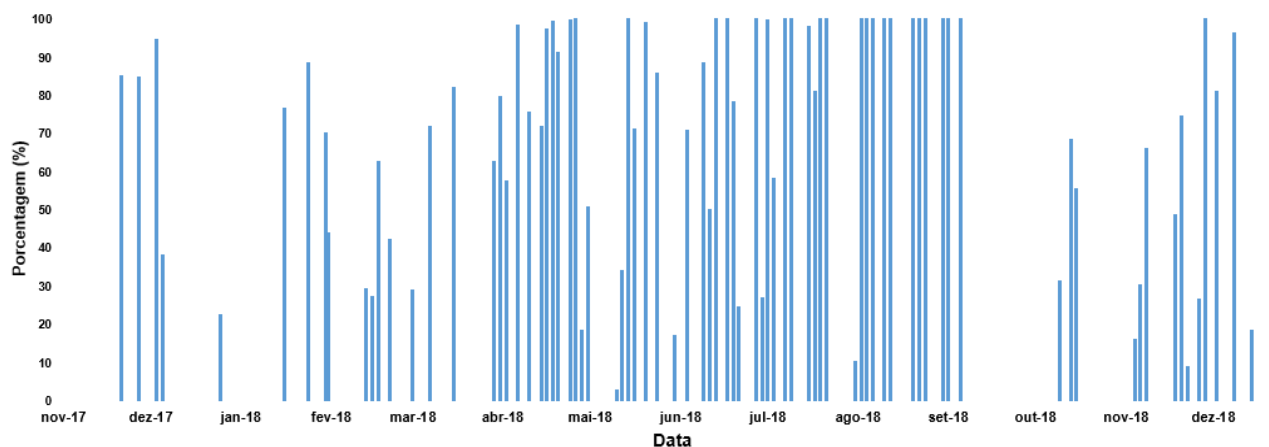


Figura 3: Porcentagem de pixels de qualidade no período de estudo novembro 17- dezembro 18

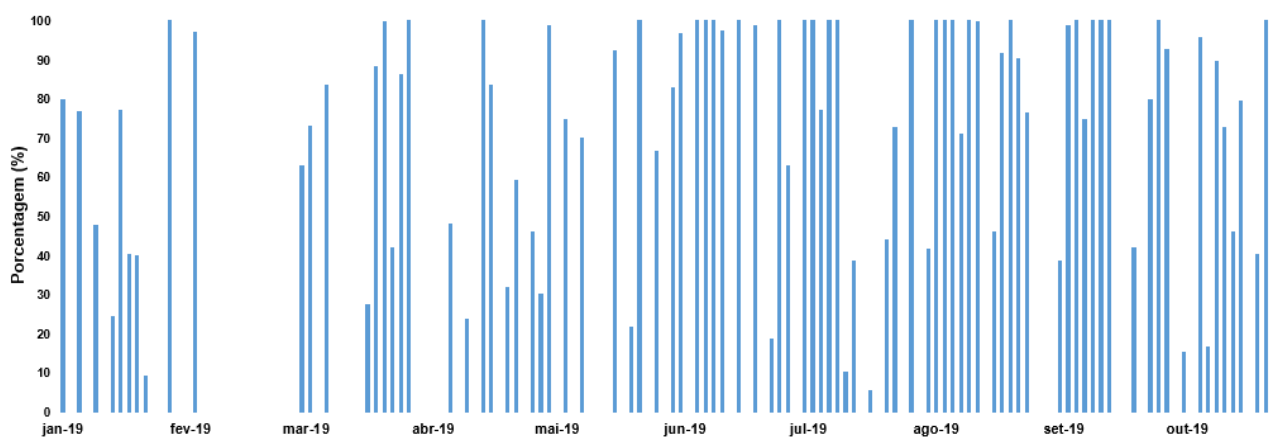


Figura 4: Porcentagem de pixels de qualidade no período de estudo janeiro 19 – outubro 19

Para a avaliação da série temporal de imagens e auxílio no processo de classificação, foi estabelecido um intervalo de 10 dias, onde foi feita a composição decendial dos dados por meio do *software Rstudio* utilizando o método da mediana do pixel. O período de estudo consiste em 24 meses, e que após esse processo totalizou 72 imagens, ou seja, 3 imagens por mês. A Figura 5 representa a composição dos dados, sendo 10 imagens faltantes. A presença de imagens faltantes acontece devido ao funcionamento satélite VEN μ S que não disponibiliza a imagem caso ela tenha em sua composição uma porcentagem de nuvem acima de 90%.

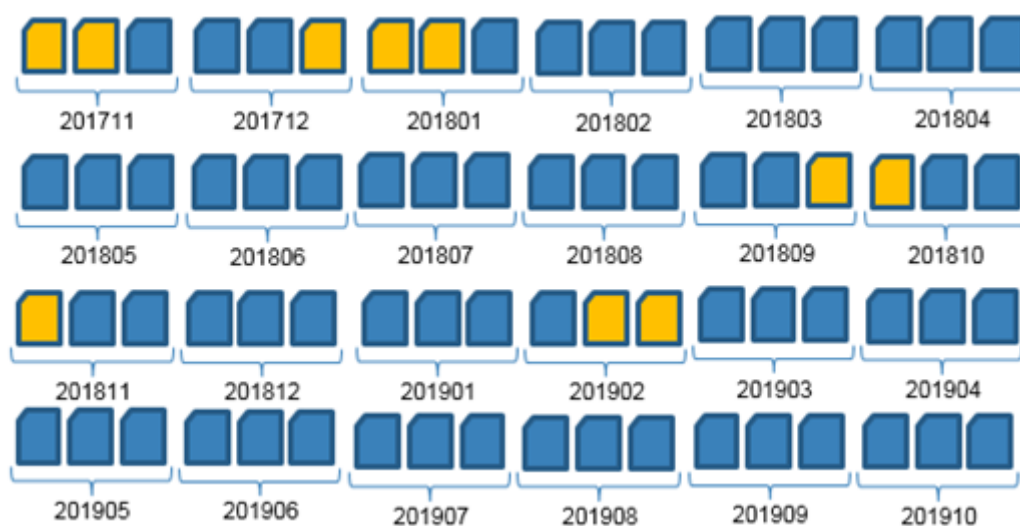


Figura 5: Representação da série temporal dos dados do satélite VEN μ S, em azul representação das imagens após a composição, e em amarelo as imagens da série que estão faltando.

3.4 Segmentação

Neste trabalho a segmentação foi feita pelo *software* QGIS. A ferramenta utilizada foi o ORFEO (*Optical and Radar Federated Earth Observation*), que apresenta um conjunto de algoritmos voltados para a segmentação. Para esta metodologia o algoritmo utilizado foi o de crescimento da região (*GenericRegionMerging*) e a imagem para segmentação foi feita utilizando todas as bandas do VENμS no dia 2 de fevereiro de 2019.

Três parâmetros foram definidos para implementação do algoritmo de segmentação da *Generic Region Merging*, que foram: o tamanho dos polígonos, o valor relacionado aos dados radiométricos e o valor referente a homogeneização espacial utilizando como entrada todas as bandas. Os parâmetros escolhidos para a segmentação, foram definidos por meio de testes baseados em análise visual, obtendo o Limiar ($t=500$), o peso para a homogeneidade espectral ($cw = 0,5$) e peso para a homogeneidade espacial ($sw = 0,5$).

A validação da segmentação foi feita por meio de comparação visual, em que a imagem segmentada foi comparada com a imagem real do VENμS. As principais formas de avaliação, foram a presença de regiões com pivô central, corpos d'água e estradas.

3.5 Índice de vegetação

Com objetivo de analisar os dados presentes na série temporal e também realizar a classificação do uso da terra, foi utilizado o índice de vegetação NDVI proposta por ROUSSE et.al (1973). Esse índice é comumente utilizado em diversos estudos, além da praticidade que permite ao usuário distinguir diferentes culturas no terreno. O NDVI foi calculado a partir das bandas 11 e 5 do satélite de acordo com a Equação 1, sendo que N representa a banda do infravermelho próximo e R a banda do vermelho.

$$NDVI = \frac{(N-R)}{(N+R)} \quad (1)$$

3.6 Preenchimento dos dados faltantes

Neste trabalho foram utilizados 5 métodos de preenchimento de dados faltantes: Interpolação ($k=1$ e $k=2$), Média, Mediana e *MissForest*. Todos os métodos são disponibilizados no *software Rstudio*.

Nos métodos de média e mediana, o algoritmo calcula um único valor que corresponde ao método utilizado, após o cálculo todos os dados faltantes são preenchidos utilizando o valor da média e da mediana da série temporal. No algoritmo dos métodos de Interpolação, o dado faltante é preenchido de acordo com os valores de seus vizinhos próximos (parâmetro k), ou seja, cada dado faltante recebe um valor diferente dos demais. Destaca-se que o pacote *imputeTS* permite diferentes métodos de cálculo através da interpolação, neste trabalho o método utilizado foi a interpolação linear.

No algoritmo *MissForest* o preenchimento é feito com base no método iterativo *Random Forest*. Esse método permite que a predição do valor faltante seja mais precisa e haja menor variabilidade entre os valores preditos. Além disso por se tratar de um método iterativo, o usuário pode definir o nome de iterações desejadas (m), desde que o critério de parada seja obedecido. O critério de parada faz com que toda a função seja interrompida quando houver aumento na variabilidade entre o valor predito (n) e o valor predito anteriormente ($n-1$), sendo n um número que varia de 1 até o número de iterações (m). Quando o critério de parada é atingido a função retorna o último valor predito e preenche o dado faltante. Neste trabalho o número máximo de iterações permitidas pelo critério de parada é três.

3.7 Validação Cruzada

Para fins de comparação foram analisadas 5 formas de preenchimento de dados em segmentos rotulados para validação da melhor forma de preenchimento para o conjunto de dados. Para a avaliação da imputação foi feita a validação cruzada *Leave-One-Out* (LOOCV) nos dados rotulados, onde apenas uma observação é usada para validação e o restante é usado para treinamento do método.

Este método de validação cruzada apresenta menos viés, já que mais observações são usadas para ajustar o modelo. Não há aleatoriedade nas divisões do conjunto de treinamento/validação. Portanto, a variabilidade é reduzida, porém o tempo de processamento aumenta visto que todos os dados são analisados. Nesse método o número de dados utilizados para treinamento é $N - 1$, sendo N o número de dados que há dados reais, ou seja, não são considerados os dados ausente da série como ilustrado no fluxograma (Figura 6).

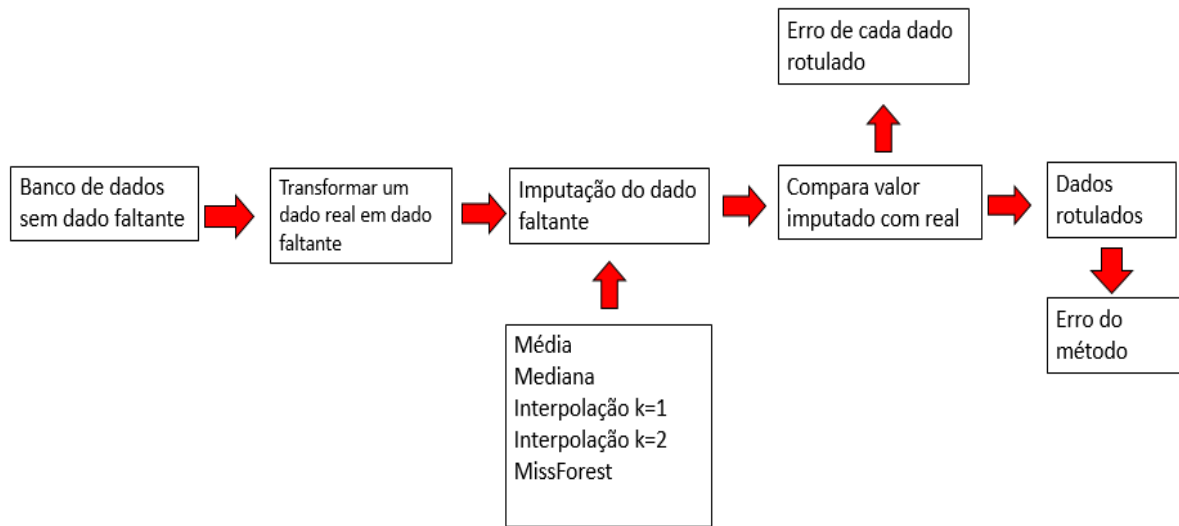


Figura 6: Fluxograma para validação cruzada

O erro é calculado pela diferença entre o dado real menos o imputado para todos os dados da série que não apresentam dados faltantes. O erro do método foi calculado pelo *Mean Absolute error* (MAE) (Equação 2) e o *Root Mean Squared Error* (RMSE) (Equação 3).

$$MAE = \left(\frac{1}{N}\right) * (\sum_{i=1}^{572} Ei) \quad (2)$$

$$RMSE = \sqrt{\left(\frac{1}{N}\right) * (\sum_{i=1}^{572} Ei)^2} \quad (3)$$

Onde N = número do banco de dados sem dado faltante (572) e Ei = (dado real - dado imputado) de cada interação

Esses erros foram calculados para as 572 séries temporais de NDVI dos dados rotulados, e por fim foi calculada a média.

3.8 Classificação do uso da terra

O trabalho de classificação foi feito através da adaptação da metodologia de classificação utilizada por SANTOS et.al (2019), sendo a adaptação necessária devido a diferenças no banco de dados que se deve principalmente pela quantidade de dados utilizados e também pelo número de índices utilizados para a classificação. Para realizar a classificação o banco de dados foi dividido em 8 classes, cada classe abrange uma boa quantidade de culturas diferentes, devido à grande quantidade de culturas presentes na área. As informações sobre cada classe podem ser vistas na Tabela 2.

Classes	Culturas	Porcentagem (%)	Quantidade de polígonos
Anual	Milho	23	130
	Trigo		
	Soja		
Cana-de-Açúcar	Cana-de-Açúcar	9	53
Lenhosa	Eucalipto	24	139
	Café		
	Pinus		
	Laranja		
Daninha*	Erva daninha	3	20
	Palha		
Nativa	Floresta nativa	10	60
	Cerrado		
Pastagem	Pasto	19	106
Água	Corpos d'água	5	30
Construção	Estradas	6	34
	Áreas urbanas		

*Classe que compreende a vegetação entre culturas

Tabela 2: Tabela referente às classes utilizadas na classificação de uso da terra e suas respectivas porcentagens em relação ao banco de dados de 572 polígonos, adaptado de SANTOS et.al 2019.

A série temporal foi limitada para um período de 6 meses, dessa forma, o classificador utilizou dados de setembro de 2018 até março de 2019. A restrição do período analisado é necessária para que o classificador possa trabalhar de forma mais eficiente e produzir melhores resultados. Além disso, o conjunto de dados apresenta grande dinamismo de rotação de culturas, o que impossibilita classificar um polígono em uma única classe durante o período total do banco de dados.

Assim, o conjunto de dados para a tarefa de classificação passa a ter um total de 572 polígonos e 18 variáveis referentes ao período escolhido (6 meses com 3 períodos). Para realizar a classificação foi feita também a divisão do banco de dados, formando os conjuntos de treino e teste. O conjunto de treino apresenta 70% do banco de dados (400 polígonos) e tem como função treinar o algoritmo de classificação e por isso necessita de maior quantidade de dados. Já o conjunto de teste tem os restantes 30% (172 polígonos) e tem como função validar o algoritmo. A divisão foi feita pela amostragem estratificada, para garantir que cada classe pudesse ser dividida na proporção de 70% e 30% para cada um dos conjuntos, evitando desigualdade na amostragem (Tabela 3)

Distribuição dos dados (70% treino - 30% teste) – Estratificada								
	Anual	Construção	Nativa	Pastagem	Cana-de-Açúcar	Água	Daninha	Lenhosas
Treino	91	24	42	74	37	21	14	97
Teste	39	10	18	32	16	9	6	42
Total	130	34	60	106	53	30	20	139

Tabela 3: Tabela referente a distribuição da amostragem estratificada para os conjuntos de treino e teste.

O algoritmo utilizado para classificação foi o *Random Forest*, método que utiliza múltiplas árvores de decisão para realizar as classificações. Os principais parâmetros necessários para o algoritmo podem ser selecionados pelo usuário, sendo eles o *mtry=7*, parâmetro relacionado ao número de variáveis e que pode ser definido utilizando-se a função *tuneRF* presente no pacote do *Random Forest* e o *ntree* que corresponde ao número de árvores criadas pelo algoritmo, os valores para *ntree* neste trabalho variam de 100, 300, 500 e 1000.

A classificação foi feita utilizando os conjuntos de dados gerados por cada método de preenchimento de dados faltantes. Dessa forma, foram 5 conjuntos diferentes a serem avaliados. A comparação foi feita através da acurácia da classificação obtida por cada um dos métodos. A Figura 7 apresenta um fluxograma para essa atividade de classificação.

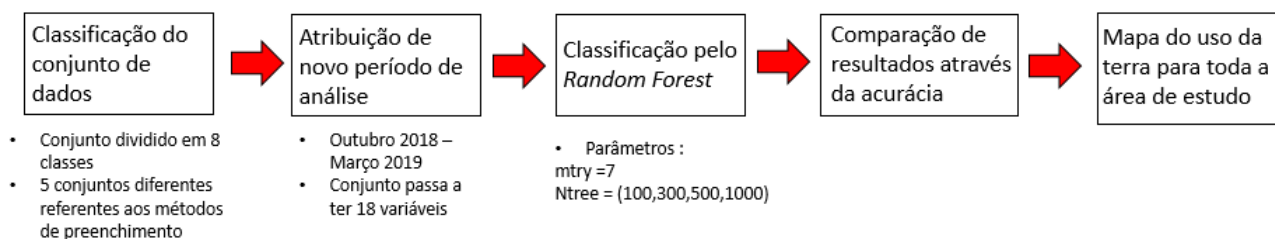


Figura 7: Fluxograma da atividade de classificação do uso da terra.

4 Resultados e Discussão

4.1 Preenchimento dos Dados Faltantes

Através do fluxograma presente na Figura 6, foi feita a validação cruzada e o cálculo do erro. A Tabela 4 apresenta o resultado dos 5 métodos de preenchimento e seus respectivos resultados.

Erro	Média	Mediana	Interpolação k=1	Interpolação k=2	Miss Forest
MAE	0,12	0,12	0,04	0,04	0,02
RMSE	0,14	0,14	0,05	0,06	0,03

Tabela 4: Valores dos erros MAE e RMSE a partir do preenchimento de dados da série temporal de imagens de NDVI.

O preenchimento feito com o *MissForest* apresentou menores erros se comparado aos demais métodos (MAE=0,02 e RMSE=0,03). A Média e a Mediana, apresentaram os maiores erros e ambas alcançaram o mesmo resultado (MAE=0,12 e RMSE=0,14). Os dois métodos de Interpolação também apresentaram resultados semelhantes, e resultados próximos ao *MissForest*. No entanto, para entender o processo de preenchimento de uma série temporal, é necessário também analisar o seu comportamento temporal da cultura ao longo do tempo para verificar se existe alguma inconsistência.

Apesar do *MissForest* ter apresentado o menor erro entre todos os métodos, as curvas de NDVI da cana-de-açúcar não foram condizentes com o comportamento padrão esperado para a cultura, o mesmo ocorreu para as curvas da Média e Mediana (Figura 8). Essa incoerência devido a forma como é feito o preenchimento, no algoritmo da Média é feito o cálculo da média de todos os pontos e todos os dados faltantes são preenchidos com esse valor, caso semelhante acontece na Mediana e no *MissForest*. Esse tipo de incoerência é mais visível numa cultura dinâmica como a cana-de-açúcar, porém numa cultura perene como a laranja esse erro não interfere a interpretação da curva (Figura 9).

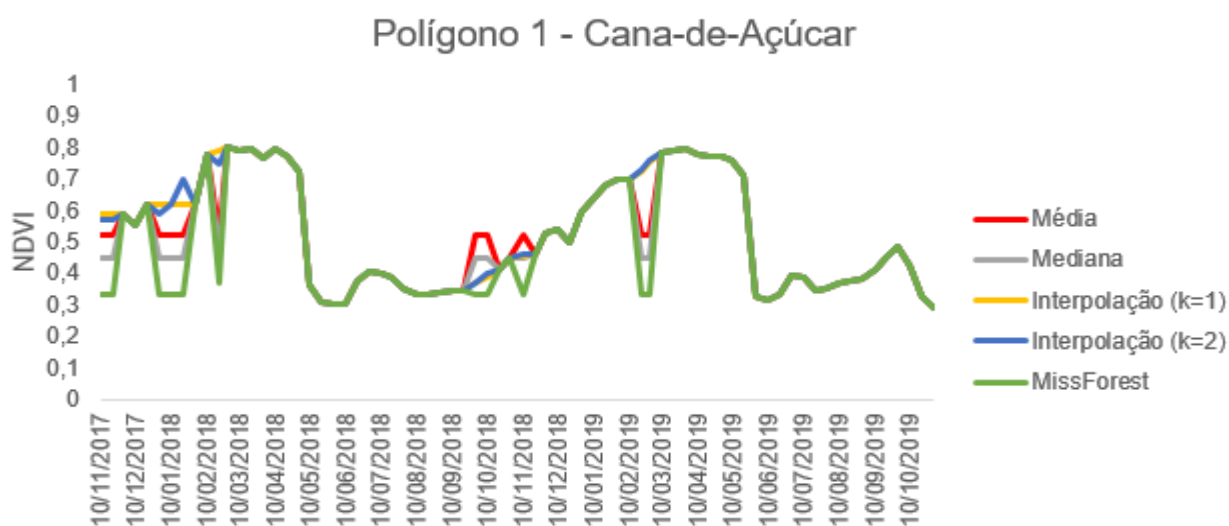


Figura 8: Figura referente a série temporal de uma cultura de Cana-de-Açúcar

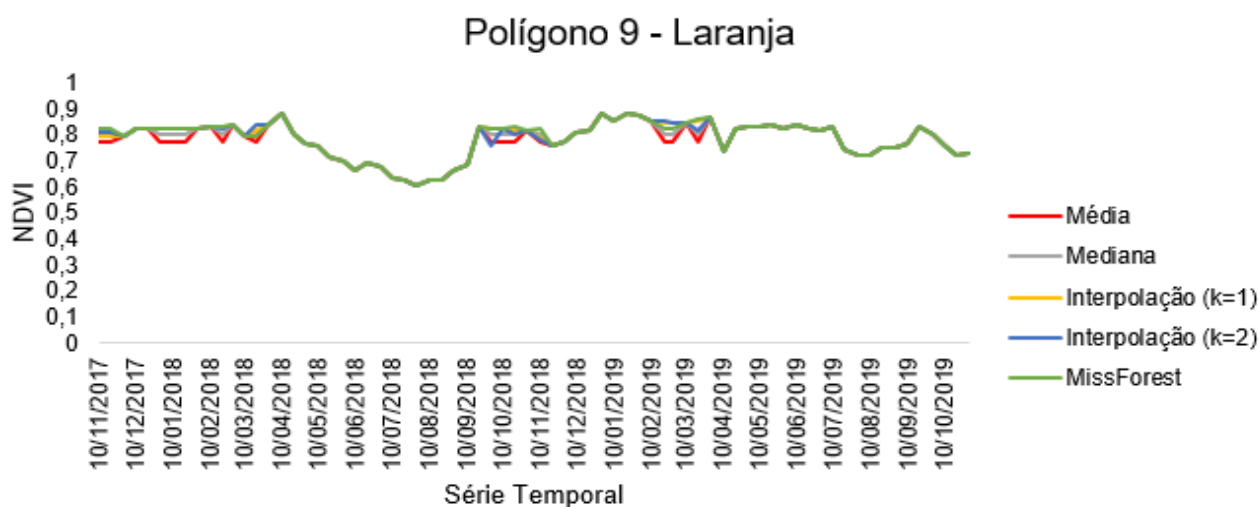


Figura 9: Figura referente a série temporal de uma cultura de laranja

Dessa forma, verificou-se que os métodos de preenchimento baseados na Média, Mediana e no *MissForest*, apresentam melhor comportamento para analisar séries temporais em que as curvas de NDVI são menos dinâmicas, ou seja, para culturas perenes como o caso da Laranja.

Vale destacar que os métodos de Interpolação conseguem se adaptar as curvas com mais precisão que os demais métodos. Essa diferença ocorre, pois, os métodos de Interpolação preenchem cada dado faltante de acordo com seus respectivos pontos adjacentes, para o $k=1$ usa-se 1 ponto antes e depois do dado faltante e para o $k=2$ a interpolação é feita com 2 pontos antes e 2 pontos após o dado faltante. Esse algoritmo de preenchimento permite que a variação da curva de NDVI ao longo do tempo, permitindo uma melhor adaptação independente da cultura analisada.

Portanto, de acordo com a análise de validação cruzada e verificação das curvas das séries temporais, pode-se dizer que os métodos de Interpolação são os mais adequados para o preenchimento dos dados faltantes deste conjunto de dados.

4.2 Classificação do uso da terra

A Tabela 5 apresenta os valores obtidos de acurácia global para cada método após a classificação utilizando o *Random Forest*. Para cada método foi testado também 4 valores da variável *ntree* (100, 300, 500 e 1000).

Acurácia da Classificação					
<i>ntree</i>	Interpolação $k=1$	Interpolação $k=2$	Média	Mediana	<i>MissForest</i>
100	0,79	0,76	0,78	0,79	0,75
300	<u>0,81</u>	0,76	0,78	0,77	0,78
500	<u>0,81</u>	0,77	0,79	0,77	0,77
1000	0,8	0,76	0,79	0,78	0,77

Tabela 5: Tabela referente aos valores da acurácia obtida pelo *Random Forest*.

O método da Interpolação ($k=1$) apresentou melhores resultados que os demais, sendo que a melhor acurácia obtida foi de 0,81. O *MissForest* que havia apresentado os menores erros de LOOCV para o preenchimento da série temporal não se mostrou eficiente no processo de classificação, obtendo em alguns momentos, o menor valor de acurácia. Assim, definiu-se o método de Interpolação ($k=1$) como o método referência para as demais etapas de análise da classificação.

SANTOS et.al (2019) obteve a Média como a melhor forma de preenchimento de dados faltantes com acurácia de 0,84, utilizando a mesma área de estudo. Além disso, SANTOS et.al (2019), também fez a classificação orientada a objeto com segmentação com parâmetros semelhantes aos utilizados neste trabalho, porém o satélite utilizado foi o Landsat, que apresenta menor resolução espacial (30m) que a do VENμS (10m).

Apesar da alta qualidade das imagens do VENμS quando comparada as do Landsat, a acurácia de Santos et.al (2019) foi melhor, isso aconteceu devido a diferenças na metodologia que são: outros índices de vegetação e as bandas espectrais, além de maior número de amostras verificadas em campo.

Na Tabela 3, observa-se que classes como a Daninha (20 polígonos) e Cana-de-Açúcar (50 polígonos) tem menor quantidade de dados que as demais classes. Trabalhar com um banco de dados maior ajuda a amenizar esse tipo de limitação, uma vez que o algoritmo tem mais informações para treinar o classificador. Como forma de avaliar o processo de classificação, foi utilizado como base os valores de omissão e comissão que podem ser vistos na Tabela 6.

Matriz de Confusão										
	Anual	Urbano	Nativa	Pastagem	Cana	Água	Daninhas	Lenhosas	Total	Comissão
Anual	37	0	0	0	2	0	1	0	40	8%
Urbano	0	8	0	1	0	0	1	0	10	20%
Nativa	0	0	12	0	0	0	0	2	14	14%
Pastagem	0	2	0	30	5	0	1	4	42	29%
Cana	2	0	0	1	9	0	0	3	15	40%
Água	0	0	0	0	0	9	0	0	9	0%
Daninhas	0	0	0	0	0	0	2	0	2	0%
Lenhosas	0	0	6	0	0	0	1	33	40	18%
Total	39	10	18	32	16	9	6	42	140	
Omissão	5%	20%	33%	6%	44%	0%	67%	21%		
Acc	0.81									

Tabela 6: Tabela referente a matriz de confusão da classificação, valores referentes ao método de Interpolação k=1. Conjunto de teste com 172 pontos (30% do conjunto total de dados).

A omissão representa a porcentagem dos dados de uma determinada classe que foram omitidas pelo classificador, sendo interpretados como pertencentes a outra classe. Já a comissão está relacionada a porcentagem que foi interpretada de forma equivocada numa determinada classe. Neste caso a classe que apresentou menores erros foi a Anual, tendo 5% dos seus dados classificados como Cana e 8% dos dados referentes a essa classe pertencentes as classes de Cana e Daninhas. A classe Água não apresenta erros, porém dado a diferença nos seus valores de NDVI para as demais classes, este era um resultado esperado cujo erro representaria uma incoerência no classificador.

Pode-se identificar também que as classes Nativa e Lenhosa apresentam erros entre si, no caso da classe nativa o seu erro de comissão é referente apenas a classe Lenhosa. Esse tipo de erro se deve pela semelhança entre as culturas presentes dentro dessas classes, tendo a presença de culturas perenes como a laranja.

A classe que apresentou os maiores erros foi a classe da Cana-de-Açúcar, tendo ambos os erros na faixa dos 40%. Isso se deve pela dinâmica deste tipo de cultura que está em constante mudança ao longo do tempo (Figura 8). Outro fator que deve ser considerado é a menor quantidade de dados nesta classe (9% de 572), o que dificulta a interpretação do comportamento dessa classe. Caso semelhante também acontece com a classe das Daninhas que apresenta apenas 3% do total de dados.

Na Figura 10, pode-se encontrar a série temporal de cada a classe ao longo do período de estudo utilizado para a classificação. Esse comportamento é analisado também fazendo uma análise estatística do conjunto de dados. Onde pode-se encontrar o maior e o menor valor de NDVI, além da média e os quartis (primeiro e terceiro) de cada data.

Analisar o comportamento das curvas ao longo do tempo auxilia a entender os erros encontrados na matriz de confusão. As classes de Lenhosa e Nativa, apresentam comportamento semelhante, porém com a diferença de que os dados da classe Nativa são mais semelhantes entre si. As semelhanças dos dados dentro da classe ajudam o classificador a encontrar um padrão, no caso da Daninha este padrão é de difícil interpretação dado ao espaçamento entre os dados, justificando o alto erro de omissão na matriz de confusão.

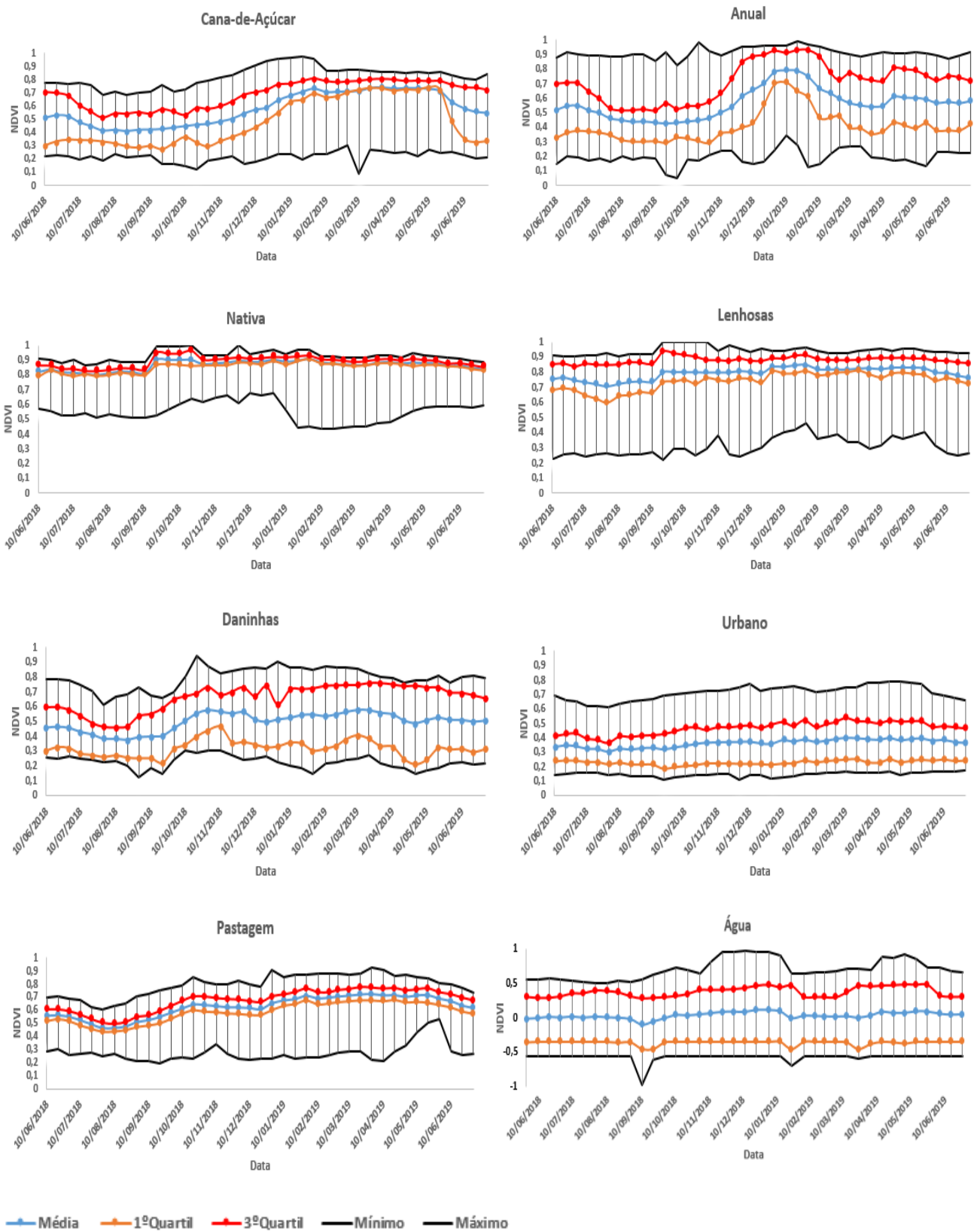


Figura 10: Figura referente ao comportamento das classes estudadas ao longo do período de estudo utilizado na classificação.

O processo de análise da classificação feita pelo *Random Forest*, teve como objetivo avaliar se o preenchimento realizado é válido e pode ser utilizada para toda a área de estudo. Com a acurácia obtida de 0,81 pode-se afirmar que o classificador teve ótimo desempenho de classificação para áreas com alta diversidade agrícola.

Foi realizado também uma classificação para toda área de estudo. A classificação foi obtida da segmentação orientada a objeto de uma imagem do VEN μ S do dia 2 de fevereiro de 2019. A imagem segmentada foi validada a partir de comparação visual entre a imagem segmentada e a imagem real disponibilizada pelo satélite. Assim foi gerado o mapa de classificação do uso da terra presente na Figura 11.

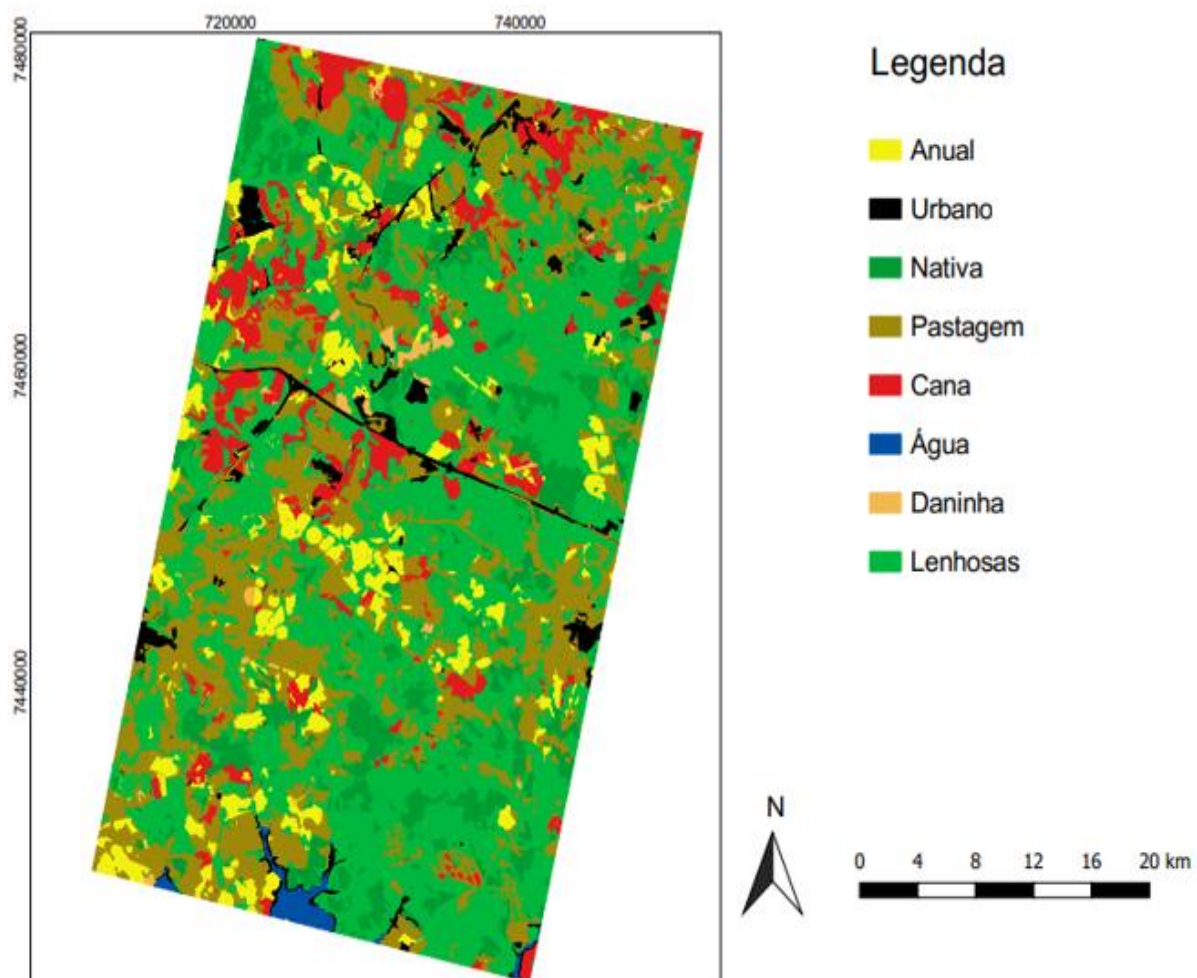


Figura 11: Figura referente ao mapa de classificação de uso da terra utilizando o método de Interpolação $k=1$ e classificação pelo algoritmo *Random Forest*.

Foi feita também a classificação para os demais métodos de preenchimento, com o objetivo de avaliar se houve diferenças nas classificações para toda área. Os mapas referentes aos demais métodos e suas respectivas matriz de confusão podem ser encontrados no apêndice desse trabalho.

Para analisar separadamente o preenchimento de cada método, foi feito uma aproximação em uma área de referência e facilitar a análise visual das imagens (Figura 12). As principais diferenças encontradas podem ser vistas nas imagens referentes aos métodos Média e Mediana, em que áreas que apresentam pivô central método de irrigação característico de culturas como milho e soja, estão classificadas como culturas Lenhosas e Nativas. O método *MissForest* também apresenta um resultado semelhante, porém com a classe Nativa.

A Figura 12 também aponta diferenças na classificação entre as classes Anual, Pastagem e Cana. Diferença que também pode ser identificada dentro da matriz de confusão. Sendo essa uma das principais limitações do classificador, o que impede acurácia maior do que 0,81.

Analisar uma área de referência ajuda a compreender possíveis erros do classificador. Porém é necessário também estender a análise para área toda. A Figura 12, também apresenta uma relação da porcentagem da área total de acordo com cada uma das 8 classes. Os valores para os dois métodos de Interpolação e o *MissForest* são semelhantes, porém os métodos da Média e Mediana apresentam diferenças significativas. A maior diferença observada é a presença de mais de 50% da classe Lenhosa. Diferença que também pode ser observada na imagem aproximada.

Observa-se também, que ao se considerar a área de estudo como um todo a diferença da área na classe Cana é de no máximo 2%, ou seja a diferença observada na área aproximada é baixa se analisada para toda área. Além disso, a porcentagem de área para as classes Cana, Água, Urbano, Anual, Daninhas e Nativa apresentam diferenças menores do que 5% validando a eficiência do classificador ao ser utilizado em toda área de estudo.

A Figura 12 apresenta diferenças entre os métodos de preenchimento. Essas diferenças também podem ser identificadas através de comparação visual entre as imagens. Destaca-se que a imagem real apresentada na Figura 12 é referente a uma data específica (2 de fevereiro de 2019), enquanto que a imagens classificadas por seus respectivos métodos, levam em conta um período de tempo de 6 meses, justificando possíveis diferenças na comparação da imagem real com as demais imagens.

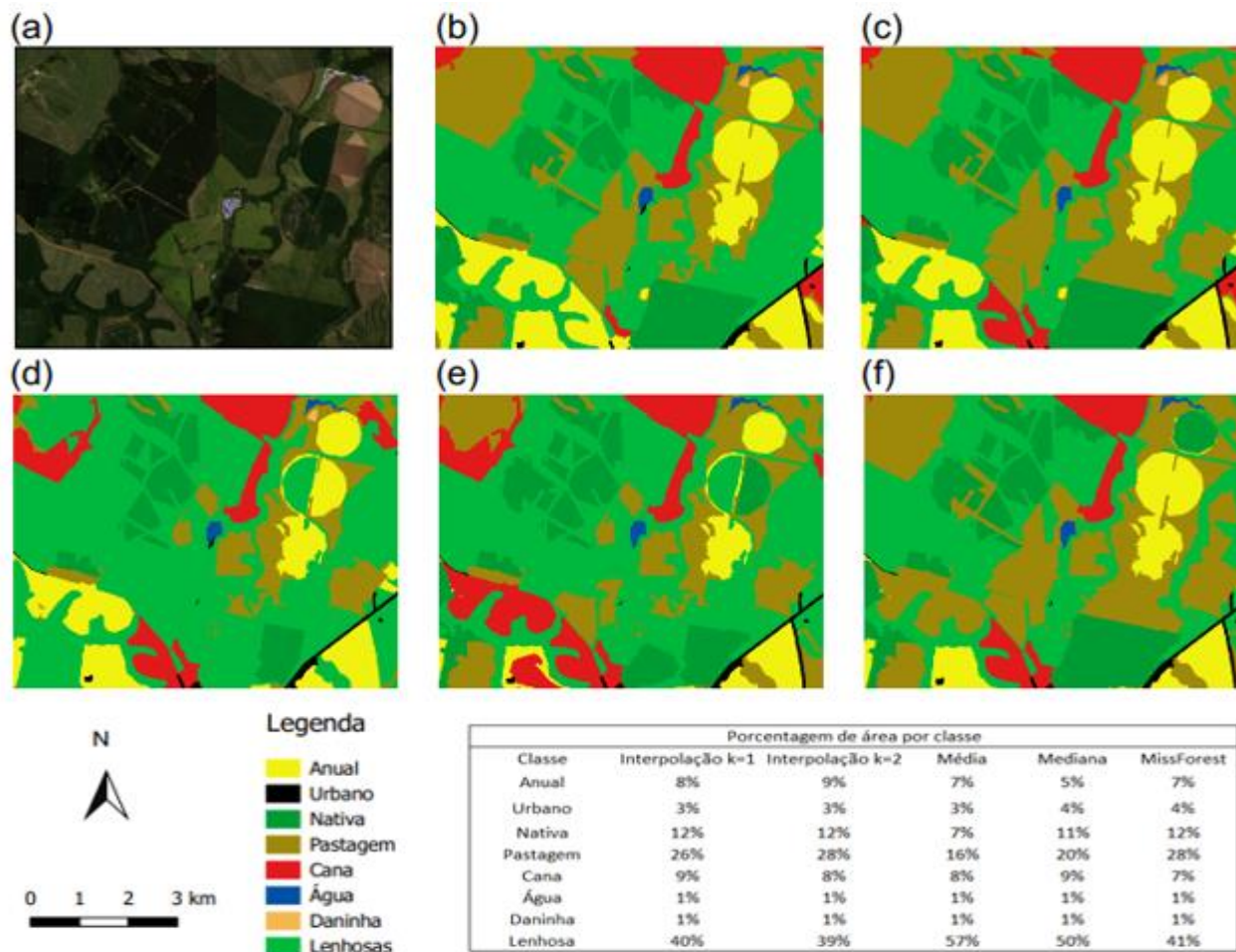


Figura 12: Aproximação dos mapas de uso da terra gerados (a) Imagem real do dia 2 de fevereiro de 2018; (b) Interpolação ($k=1$); (c) Interpolação ($k=2$); (d) Média; (e) Mediana; (f) *MissForest*. Composição RGB 5-4-3, e relação da porcentagem de área total para as 8 classes.

5 Conclusão

Esse trabalho comparou diferentes métodos de preenchimentos de dados faltantes em séries temporais de imagens de NDVI e posteriormente testamos essas séries para a classificação de uso da terra através do algoritmo *Random Forest*. Através dos resultados obtidos e as análises realizadas pode-se concluir que o método de preenchimento de dados faltantes que apresentou melhores resultado, foi o método de Interpolação ($k=1$). Apresentando um dos menores erros ($MAE=0,04$). Destaca-se também a importância de analisar o comportamento da série temporal, pois apenas a análise estatística não é capaz de identificar possíveis incoerências no preenchimento da série. O método de Interpolação ($k=1$) se mostrou capaz de se adaptar a todas as culturas estudadas, enquanto os métodos como Média e Mediana se comportam melhor para culturas perenes. Demonstrando a importância de se fazer a análise da série temporal por meio de análise visual e estatística.

Em relação ao classificador *Random Forest*, a Interpolação ($k=1$) também apresentou resultados coerentes com o esperado e uma acurácia de 0,81, embora algumas classes tenham apresentado altos erros de omissão e comissão. Esses erros foram atribuídos aos principalmente devido à quantidade de amostras de treinamento para essas classes, além disso, alguns erros foram verificados devido a semelhança de algumas classes (Nativa e Lenhosas). Apesar das diferenças, todos os métodos quando aplicados a toda área de estudo, tiveram bons resultados.

A presença de nuvens é considerada como uma dificuldade para a aplicação de mapas de classificação do uso da terra no Brasil. Este trabalho mostra que esse tipo de problema pode ser amenizado dado que o preenchimento de dados faltantes torna possível a elaboração de um mapa próximo a realidade.

6 Referências Bibliográficas

- ANTUNES, José L.F. **Uso da análise de séries temporais em estudos epidemiológicos**. Epidemiol. Serv. Saúde, 2015.
- ARÉVALO, P.; OLOFSSON, P.; WOODCOCK, C. E. **Continuous monitoring of land change activities and post-disturbance dynamics from Landsat time series: A test methodology for REDD+ reporting**. *Remote Sensing of Environment*, v. 238, n. January 2018, p. 111051, 2020.
- A.K. Bhandari, A. Kumar, **“Feature Extraction using Normalized Difference Vegetation Index (NDVI): A Case Study of Jabalpur City”**, *Proceedings of Communication, Computing & Security*. Procedia Technology Volume 6, pp. 612– 621, 2012
- BAYER, F. M.; SOUZA, A. M. **Wavelets e modelos tradicionais de previsão: Um estudo comparativo**. *Revista Brasileira de Biometria*, São Paulo. v.28, p.40-61, 2010.
- BELGIU, M.; CSILLIK, O. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sensing of Environment*, v. 204, n. October 2017, p. 509–523, 2018.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32
- BORGES, E.F *et al.* **SÉRIES TEMPORAIS DE EVI DO MODIS PARA O MAPEAMENTO DE USO E COBERTURA VEGETAL DO OESTE DA BAHIA**. Embrapa, 2014
- CARVALHO, Melissa M. **Dados faltantes em análises: uma revisão sobre métodos estatísticos flexíveis a incompletude**. UFPR, 2017. Disponível em: eventos.ufpr.br/smne/SMNE2017. Acesso em: 24 mar. 2020.
- CHINO, D. Y. T.; ROMANI, L. A. S.; TRAINA, A. J. M. Chino, Romani e Traina 2010 - **Construindo Séries Temporais de Imagens de Satélite para Sumarização de Dados Climáticos e Monitoramento de Safras Agrícolas**.pdf. [s.d.].
- DIAO, C. Complex network-based time series remote sensing model in monitoring the fall foliage transition date for peak coloration. *Remote Sensing of Environment*, v. 229, n. May, p. 179–192, 2019
- ESQUERDO, Júlio César D.M. **O uso de séries temporais no monitoramento agrícola e ambiental**. Embrapa, Embrapa Informática Agropecuária, 2016.

GASPARINI, K. A. C. et al. **Técnicas de geoprocessamento e sensoriamento remoto aplicadas na identificação de conflitos do uso da terra em seropédica-RJ**. Floresta e Ambiente, v. 20, n. 3, p. 296–306, 2013.

HAGOLLE, O. et al. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, LandSat, VENμS and Sentinel-2 images. **Remote Sensing**, v. 7, n. 3, p. 2668–2691, 2015.

HAREL, Ofer; ZHOU, Xiao-Hua. Multiple imputation: review of theory, **implementation and software**. **Statistics in medicine**, v. 26, n. 16, p. 3057-3077, 2007.

HERSCOVITZ, J. et al. **Venμs-Mission Enhancement Using Electric Propulsion : First In-Flight Results of the EPS and Technological Mission SPACE PROPULSION 2018 Venμs – Mission Enhancement Using Electric Propulsion : First In-Flight Results of the EPS and Technological Mission**. n. May, 2018.

HOLLOWAY, J. et al. **A Decision Tree Approach for Spatially Interpolating Missing Land Cover Data and Classifying Satellite Images**. **Remote Sensing**, v. 11, n. 15, p. 1796, 2019.

LIAO, C. et al. Synergistic use of multi-temporal RADARSAT-2 and VENμS data for crop classification based on 1D convolutional neural network. **Remote Sensing**, v. 12, n. 5, p. 1–17, 2020.

LUCIANO, A.C.; Picoli, M.C.A.; Rocha, J.V.; Franco, H.; Junqueira, C.; Sanches, G.M. **Generalized space-time classifiers for monitoring sugarcane areas in Brazil**. **Remote Sens. Environ.** 2018, 215, 438–451, doi:10.1016/j.rse.2018.06.017.

JAMES, G.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: With Applications in R**, Springer, 2013

MANABE, V. D.; MELO, M. R. S.; ROCHA, J. V. Framework for mapping integrated crop-livestock systems in Mato Grosso, Brazil. **Remote Sensing**, v. 10, n. 9, 2018.

MARTINELLI, M. **Clima do Estado de São Paulo**. Confins Online, 2010. Disponível em: <http://journals.openedition.org/confins/6348>. Acesso em Março de 2020

MORITZ, Steffen. **Time Series Missing Value Imputation**. [S. l.], 1 jul. 2017. Disponível em: <https://cran.r-project.org/web/packages/imputeTS/imputeTS.pdf>. Acesso em: 4 maio 2020.

NUNES, Luciana N *et al.* **Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos.** Cad. Saúde Pública, [s. l.], v. 25, ed. 2, 1 fev. 2009.

POGGIO, L.; GIMONA, A.; BROWN, I. Spatio-temporal MODIS EVI gap filling under cloud cover: An example in Scotland. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 72, p. 56–72, 2012

POVH, F. P. *et al.* **Comportamento do NDVI obtido por sensor ótico ativo em cereais.** Pesquisa Agropecuária Brasileira, v. 43, n. 8, p. 1075–1083, 2008.

PRATAMA, I. *et al.* A review of missing values handling methods on time-series data. **2016 International Conference on Information Technology Systems and Innovation, ICITSI 2016 - Proceedings**, n. October, 2017.

R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. R Development Core Team (2020).

ROLIM, G. D. S.; APARECIDO, L. E. D. O. Camargo , **Köppen and Thornthwaite climate classification systems in defining climatical regions of the state of São Paulo** , Brazil. n. June, 2015.

ROUSE, J. W. *et al.* **Monitoring vegetation systems in the great plains with ERTS.** In: **EARTH RESOURCES TECHNOLOGY SATELLITE-1 SYMPOSIUM**, 3., 1973, Washington. Proceedings... Washington: NASA, v.1, p.309-317, 1973

Rudorff, B.F.T., de Aguiar, D.A., da Silva, W.F., Sugawara, L.M., Adami, M., Moreira, M.A., 2010. **Studies on the rapid expansion of sugarcane for ethanol production in Sao Paulo state (Brazil) using Landsat data.** Remote Sens. 2, 1057–1076

SANTOS, C. L. M. DE O. *et al.* **Classification of crops, pastures, and tree plantations along the season with multi-sensor image time series in a subtropical agricultural region.** Remote Sensing, v. 11, n. 3, p. 1–27, 2019.

SHIMABUKURO, Y. E.; ABÍLIO, O.; JÚNIOR, D. C. Espaço & Geografia, Vol.6, N. v. 6, 2003.

SHIRATSUCHI, Luciano S *et al.* **Sensoriamento Remoto: conceitos básicos e aplicações na Agricultura de Precisão.** Embrapa, 2014.

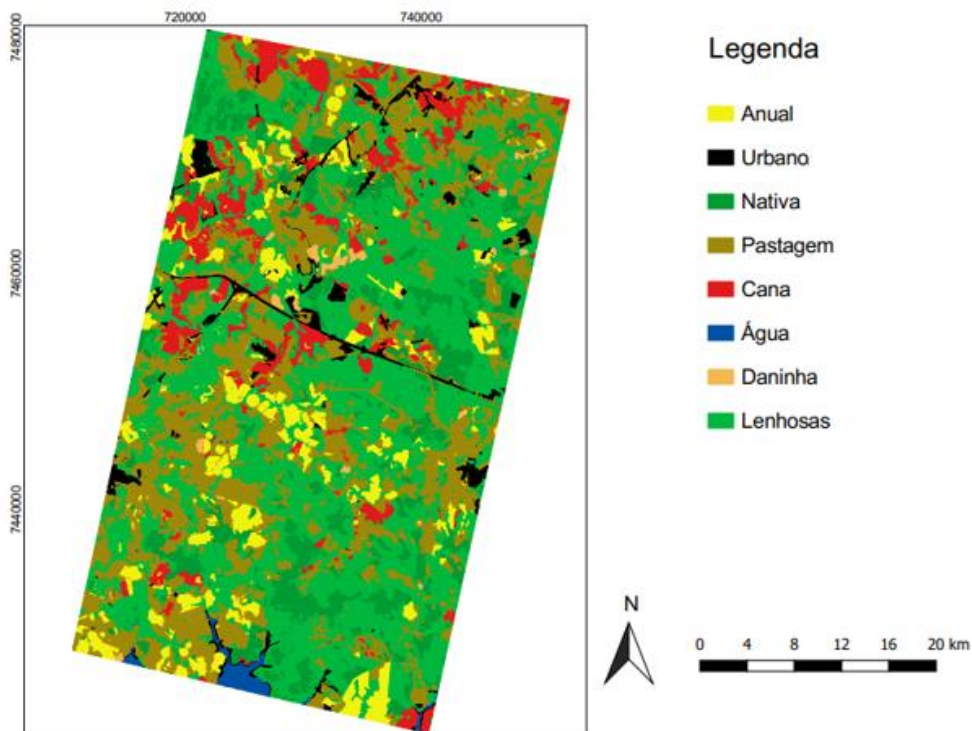
STEKHOVEN, Daniel J. **Nonparametric Missing Value Imputation using Random Forest.** [S. l.], 31 dez. 2013. Disponível em: <https://cran.r-project.org/web/packages/missForest/missForest.pdf>. Acesso em: 4 maio 2020.

WEISS, Daniel J *et al.* An effective approach for gap-filling continental scale remotely sensed time-series. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], 6 nov. 2014.

WERNER, J. P. S.; OLIVEIRA, S. R. D. M.; ESQUERDO, J. C. D. M. Mapping cotton fields using data mining and MODIS time-series. **International Journal of Remote Sensing**, v. 41, n. 7, p. 2457–2476, 2020.

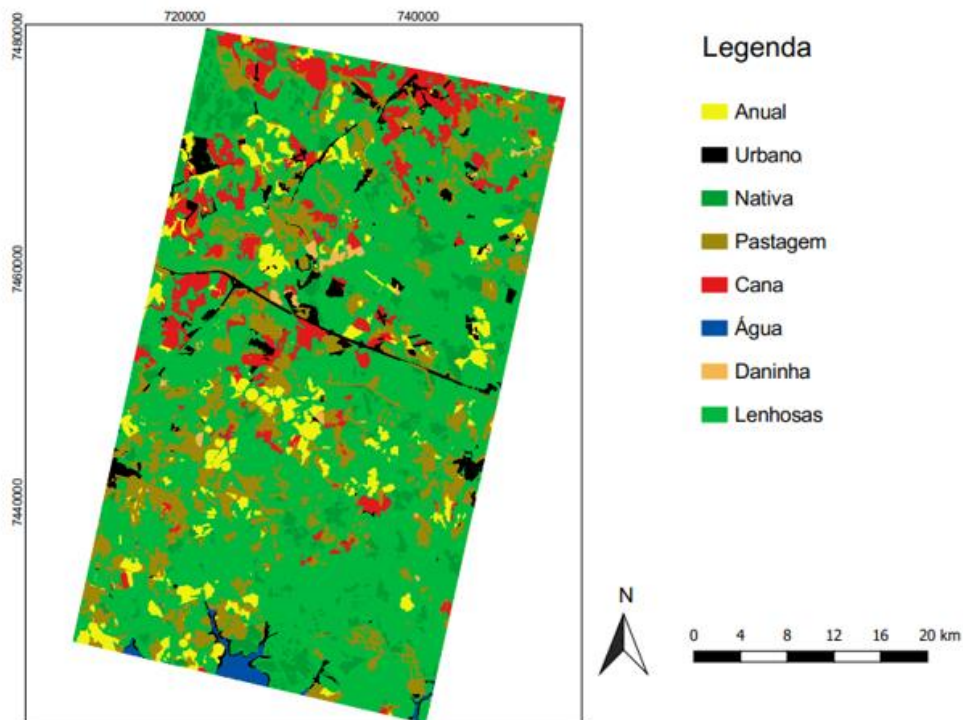
Apêndice:

A: Mapa de classificação do uso do solo e matriz de confusão, método Interpolação k=2



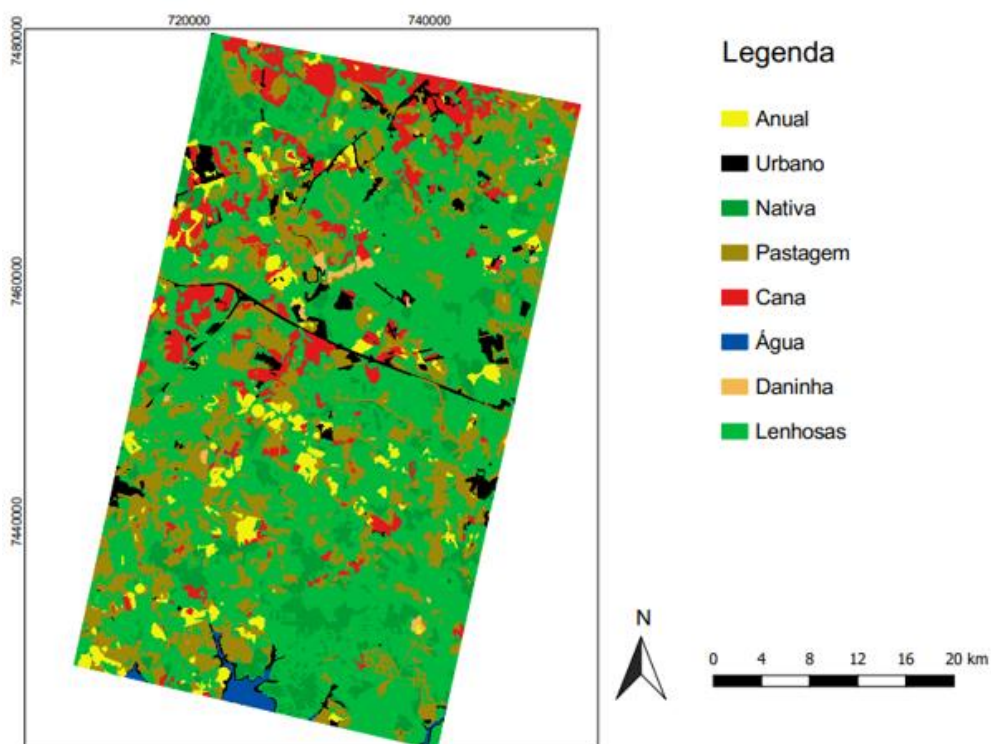
Matriz de Confusão										
	Anual	Urbano	Nativa	Pastagem	Cana	Água	Daninhas	Lenhosas	Total	Comissão
Anual	35	0	0	1	4	0	2	3	45	22%
Urbano	0	10	0	0	1	0	0	0	11	9%
Nativa	0	0	11	0	0	0	0	2	13	15%
Pastagem	2	0	0	24	1	0	0	4	31	23%
Cana	2	0	0	4	9	0	2	3	20	55%
Água	0	0	0	0	0	9	0	0	9	0%
Daninhas	0	0	0	1	0	0	2	0	3	33%
Lenhosas	0	0	7	2	1	0	0	30	40	25%
Total	39	10	18	32	16	9	6	42	130	
Omissão	10%	0%	39%	25%	44%	0%	67%	29%		
Acc	0,76									

B: Mapa de classificação do uso do solo e matriz de confusão, método Média



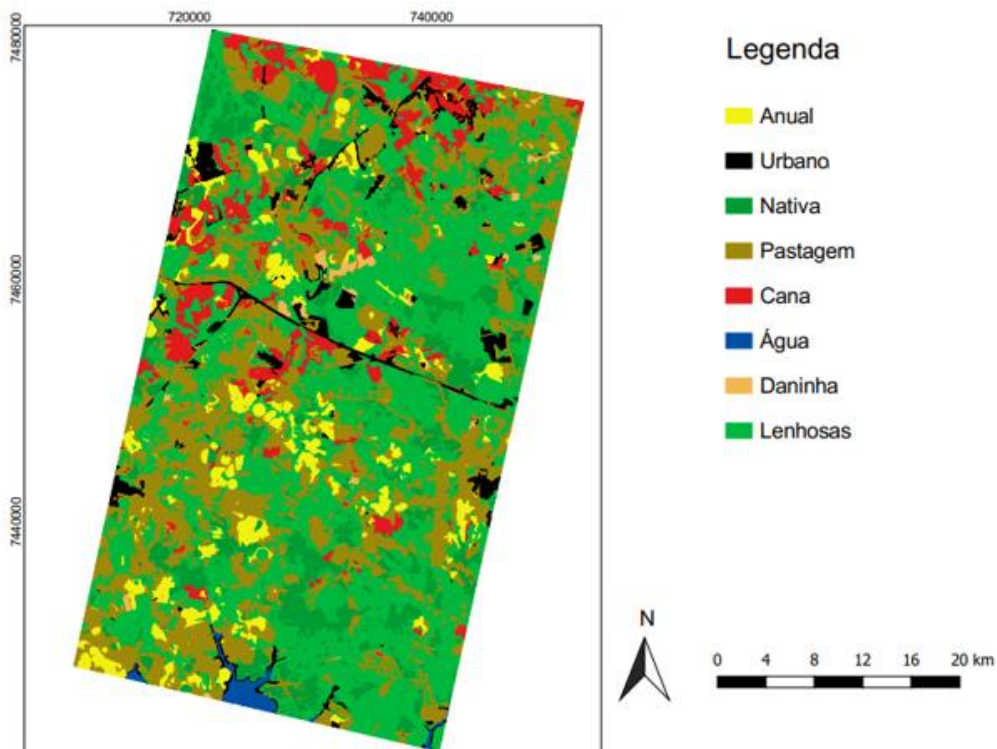
Matriz de Confusão										
	Anual	Urbano	Nativa	Pastagem	Cana	Água	Daninhas	Lenhosas	Total	Comissão
Anual	35	0	0	4	1	0	1	0	41	15%
Urbano	0	7	0	0	0	0	0	0	7	0%
Nativa	0	0	12	0	0	0	0	4	16	25%
Pastagem	1	3	0	25	5	0	1	4	39	36%
Cana	3	0	0	2	10	0	1	0	16	38%
Água	0	0	0	0	0	9	0	0	9	0%
Daninhas	0	0	0	0	0	0	3	0	3	0%
Lenhosas	0	0	6	1	0	0	0	34	41	17%
Total	39	10	18	32	16	9	6	42	135	
Omissão	10%	30%	33%	22%	38%	0%	50%	19%		
Acc	0,78									

C: Mapa de classificação do uso do solo e matriz de confusão, método Mediana



Matriz de Confusão										
	Anual	Urbano	Nativa	Pastagem	Cana	Água	Daninhas	Lenhosas	Total	Comissão
Anual	35	0	0	1	4	0	2	1	43	19%
Urbano	0	10	0	0	1	0	0	0	11	9%
Nativa	0	0	12	0	0	0	0	3	15	20%
Pastagem	2	0	0	26	1	0	0	5	34	24%
Cana	2	0	0	2	7	0	2	2	15	53%
Água	0	0	0	0	0	9	0	0	9	0%
Daninhas	0	0	0	1	0	0	2	0	3	33%
Lenhosas	0	0	6	2	3	0	0	31	42	26%
Total	39	10	18	32	16	9	6	42	132	
Omissão	10%	0%	33%	19%	56%	0%	67%	26%		
Acc	0,77									

D: Mapa de classificação do uso do solo e matriz de confusão, método *MissForest*



Matriz de Confusão										
	Anual	Urbano	Nativa	Pastagem	Cana	Água	Daninhas	Lenhosas	Total	Comissão
Anual	35	0	0	1	3	0	1	1	41	15%
Urbano	0	10	0	0	1	0	0	0	11	9%
Nativa	0	0	11	0	0	0	0	2	13	15%
Pastagem	0	0	0	24	1	0	1	4	30	20%
Cana	3	0	0	1	10	0	2	2	18	44%
Água	0	0	0	0	0	9	0	0	9	0%
Daninhas	0	0	0	1	0	0	2	0	3	33%
Lenhosas	1	0	7	5	1	0	0	33	47	30%
Total	39	10	18	32	16	9	6	42	134	
Omissão	10%	0%	39%	25%	38%	0%	67%	21%		
Acc	0,78									