UILSON RICARDO VENÂNCIO AIRES

MODELAGEM TEMPORAL E ESPACIAL DA CONCENTRAÇÃO SUPERFICIAL DE SEDIMENTOS UTILIZANDO SENSORIAMENTO REMOTO ORBITAL E APRENDIZADO DE MÁQUINA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

Orientador: Demetrius David da Silva
Coorientadores: Elpídio Inácio Fernandes Filho
Lineu Neiva Rodrigues

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

A298m
2022

Aires, Uilson Ricardo Venâncio, 1989-
Modelagem temporal e espacial da concentração superficial de sedimentos utilizando sensoriamento remoto orbital e aprendizado de máquina / Uilson Ricardo Venâncio Aires. - Viçosa, MG, 2022.
1 tese eletrônica (114 f.): il.

Inclui anexos.
Inclui apêndices.
Orientador: Demetrius David da Silva.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Engenharia Agrícola, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.160
Modo de acesso: World Wide Web.

1. Sensoriamento remoto. 2. Satélites artificiais - Sensoriamento remoto. 3. Aprendizado do computador. 4. Sedimentos fluviais. 5. Barragens de rejeitos - Doce, Rio, Bacia (MG e ES). I. Silva, Demetrius David da, 1966-. II. Universidade Federal de Viçosa. Departamento de Engenharia Agrícola. Programa de Pós-Graduação em Engenharia Agrícola. III. Título.

CDD 22. ed. 621.3678

Bibliotecário(a) responsável: Renata de Fátima Alves CRB6/2578

UILSON RICARDO VENÂNCIO AIRES

MODELAGEM TEMPORAL E ESPACIAL DA CONCENTRAÇÃO SUPERFICIAL DE SEDIMENTOS UTILIZANDO SENSORIAMENTO REMOTO ORBITAL E APRENDIZADO DE MÁQUINA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Doctor Scientiae*.

APROVADA: 14 de março de 2022.

Assentimento:

Uilson Ricardo Venâncio Aires
Autor

Demetrius David da Silva Orientador
Orientador

À minha família pelo apoio e incentivo para continuar em meus estudos, dedico.

AGRADECIMENTOS

À Universidade Federal de Viçosa e ao Departamento de Engenharia Agrícola, pela oportunidade de aperfeiçoamento da minha formação acadêmica.

Ao professor Demetrius David da Silva, pela confiança e orientação ao longo do curso de doutorado e pelo valioso direcionamento para a realização desta pesquisa.

Aos professores e coorientadores Elpídio Inácio Fernandes Filho e Lineu Neiva Rodrigues, pelas ideias, críticas e sugestões no desenvolvimento deste trabalho.

Ao professor Eduardo Morgan Uliana pelo tempo disponibilizado e sugestões fundamentais para o desenvolvimento desta pesquisa.

Aos professores Celso e Ricardo pelas recomendações e questionamentos que enriqueceram a qualidade deste trabalho.

À minha família pelo incentivo na continuação da minha formação acadêmica e compreensão pelas horas em que estive ausente para a realização deste trabalho.

À Jasmine pelo carinho e companheirismo que fizeram as longas horas de pesquisa serem menos cansativa.

Aos amigos de pós-graduação do Centro de Referências em Recursos Hídricos, pela convivência e por dividirem as dificuldades e conquistas.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, e com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq) – Processo 141059/2018-4.

E a todos que que de alguma forma contribuíram para o desenvolvimento desta pesquisa.

Muito obrigado!

BIOGRAFIA

Uilson Ricardo Venâncio Aires, filho de Ilson Benedito Venâncio Aires e Elisabete de Oliveira Venâncio Aires, nasceu em Angatuba, SP, em 10 de julho de 1989.

Em março de 2007, ingressou-se no curso de Agronegócios na Faculdade de Tecnologia de São Paulo (FATEC), em Itapetininga, SP, concluindo-o em dezembro de 2009.

Em março de 2011, ingressou-se no curso de Engenharia Agrícola pela Universidade Federal de Lavras (UFLA), em Lavras, MG, concluindo-o em fevereiro de 2016. Foi bolsista de Iniciação Científica por dois anos pelo programa PIBIC/FAPEMIG no laboratório de hidrologia florestal do departamento de Engenharia de Água e Solo da UFLA.

Em fevereiro de 2016, ingressou-se no Programa de Pós-graduação, em Nível de Mestrado, no Departamento de Engenharia Agrícola da Universidade Federal de Viçosa (UFV), submetendo-se à defesa de Dissertação em fevereiro de 2018.

Em março de 2018, ingressou-se no Programa de Pós-graduação, em Nível de Doutorado, no Departamento de Engenharia Agrícola da Universidade Federal de Viçosa (UFV), com período de Doutorado Sanduiche na *University of Florida* entre setembro de 2019 a fevereiro de 2020, submetendo-se à defesa de Tese em março de 2022.

RESUMO

AIRES, Uilson Ricardo Venâncio, D.Sc., Universidade Federal de Viçosa, março de 2022. **Modelagem temporal e espacial da concentração superficial de sedimentos utilizando sensoriamento remoto orbital e aprendizado de máquina.** Orientador: Demetrius David da Silva. Coorientadores: Elpídio Inácio Fernandes Filho e Lineu Neiva Rodrigues.

Entender a dinâmica da produção, transporte e deposição de sedimentos é de interesse em diversas áreas do conhecimento, pois a concentração de sedimentos em corpos hídricos é a principal causa de problemas relacionados com a qualidade da água, assoreamento de rios e reservatórios. No entanto, a medição em campo desta informação é bastante trabalhosa, o que dificulta a obtenção de bases de dados detalhadas e contínuas. Desta forma, o objetivo deste trabalho é modelar a variação temporal e espacial da concentração superficial de sedimentos (CSS) na bacia hidrográfica do rio Doce utilizando sensoriamento remoto orbital e modelos de aprendizado de máquina. A modelagem da CSS foi realizada a partir de duas metodologias distintas. A primeira abordou a utilização de sensoriamento remoto, em que foram utilizadas imagens orbitais dos sensores da contestação *MultiSpectral Instrument* (MSI)/Sentinel 2 e *Operational Land Imager/*(OLI) Landsat 8. Estabeleceu-se relação entre a refletância estimada pelos satélites com a CSS observada em campo medidas pela Agência Nacional de Águas e Saneamento (ANA) e pela Fundação Renova, por meio dos modelos regressão linear simples e múltipla (RLS e RLM), *least absolute shrinkage and selection operator* (LASSO) e e*lastic net*. A segunda metodologia abordou a aplicação de modelos baseados em aprendizado de máquina, em que se utilizou dados históricos da medição da CSS realizada pela ANA em sete estações sedimentométricas instaladas ao longo da calha do rio Doce. Para a predição CSS foram utilizadas 62 variáveis preditoras derivadas das informações de declividade, pedologia, uso e cobertura da terra, precipitação, vazão fluvial, velocidade fluvial, evapotranspiração real, escoamento superficial, umidade do solo, temperatura e *normalized difference vegetation index*. Os seguintes algoritmos de aprendizado de máquina foram utilizados: r*andom forest (RF)*, c*ubist, support vector machines (SVMs)*, *extreme gradient boosting machine (XGboost)* e regressão LASSO. Nas duas metodologias foram utilizadas a validação cruzada *leave-one-out* para o treinamento e testes dos modelos. As métricas adotadas para avaliação de desempenho foram erro médio absoluto (MAE), raiz do erro médio quadrático (RMSE), porcentagem do viés (PBIAS), coeficiente de Nash–Sutcliffe (NSE), índice de concordância de Willmot (d), coeficiente de determinação (R^2), coeficiente de Kling-Gupta (KGE) e índice de eficiência (c).

A banda do infravermelho próximo apresentou forte relação linear com a CSS, tanto utilizando o satélite MSI/Sentinel 2, quanto o OLI/Landsat 8. Dentre os modelos de regressão linear que utilizam múltiplas variáveis, a regressão linear múltipla, as regressões LASSO e *Elastic Net* apresentaram bom desempenho para a predição da CSS. Entretanto, a regressão LASSO e *Elastitic Net* facilitam na definição do conjunto ótimo de variáveis. Os mapas de fluxos de sedimentos indicam redução da CSS na calha do rio Doce em anos mais recentes, o que pode ser indicativo de que parte do material oriundo do rompimento da barragem de rejeitos de Fundão, em 2015, pode ter sido carreado pelos processos de ressuspensão e transporte de sedimentos. Bons resultados foram obtidos com a utilização de algoritmos de aprendizado de máquina para a predição da CSS na bacia hidrográfica do rio Doce, com destaque para os modelos cubist e XGBoost, que apresentaram o menor erro de predição e métricas de eficiência mais elevadas. As variáveis mais importantes para os modelos de predição se configuraram nas vazões fluviais diárias da data da coleta dos sedimentos e as vazões defasadas no tempo. A precipitação média diária acumulada também foi importante na modelagem dos sedimentos. A utilização dos modelos de aprendizado de máquina pode ser de grande auxílio para o monitoramento dos sedimentos, e servir como ferramenta para entender a dinâmica da produção de sedimentos na bacia hidrográfica do rio do Doce ao longo do tempo.

Palavras-chave: Sentinel 2 e Landsat 8. Modelagem hidrossedimentológica. Aprendizado supervisionado. Barragem de rejeitos de Fundão.

ABSTRACT

AIRES, Uilson Ricardo Venâncio, D.Sc., Universidade Federal de Viçosa, March, 2022. **Temporal and spatial modeling of suspended sediment concentration using orbital remote sensing and machine learning.** Adviser: Demetrius David da Silva. Co-advisers: Elpídio Inácio Fernandes Filho and Lineu Neiva Rodrigues.

The processes of sediment yield, transport, and deposition are fundamental for several areas of knowledge because sediment concentration causes problems related to water quality and silting of rivers and reservoirs. However, field measurement of sediment concentration is quite laborious, and therefore, obtaining detailed and continuous databases of this information is difficult. Thus, the aim of this study was to model the temporal and spatial variation of the superficial sediment concentration (SSC) in the Doce River Basin using orbital remote sensing and machine learning models. We used two different methodologies to predict SSC. The first included the application of remote sensing information, wherein we used orbital images from the multispectral instrument (MSI)/Sentinel 2 and the operational land imager (OLI)/Landsat 8. Simple and multiple linear regression models (SLR and MLR), least absolute shrinkage and selection operator (LASSO) regression, and elastic net regression were used to establish a relationship between the reflectance estimated by the satellites and the SSC measured in the field by the *Agência Nacional de Águas e Saneamento Básico* (ANA) and the Renova foundation. The second methodology involved the application of machine learning models using historical data of the SSC measurements monitored by the ANA at seven sediment gauge stations installed along the channel of the Doce River. Sixty-two predictor variables were used for SSC prediction derived from the following information: slope, pedology, land use and cover, precipitation, river streamflow, river velocity, actual evapotranspiration, surface runoff, soil moisture, temperature, and normalized difference vegetation index. The following machine-learning algorithms were used: random forest, cubist, support vector machines, extreme gradient boosting machine (XGBoost), and LASSO regression. In both methodologies, leave-one-out cross-validation was used to train and test the models. The metrics for performance evaluation were the mean absolute error, root mean square error, percentage of bias, Nash–Sutcliffe coefficient, Willmot agreement index, coefficient of determination), Kling–Gupta coefficient, and efficiency index. Using the MSI/Sentinel 2 and OLI/Landsat 8 satellites, the near-infrared band showed a strong linear relationship with SSC. Among the linear regression models that use multiple variables, both the MLR and LASSO and elastic net regressions

performed well for SSC prediction. However, LASSO and elastic net best defined the optimal set of variables. The sediment flux maps indicated a reduction in SCC in the Doce River in recent years, which may indicate that part of the material from the rupture of the Fundão tailings dam may have been transported by resuspension and transport of sediment. Good results were obtained with the machine learning algorithms for SSC prediction in the Doce River Basin. The cubist and XGBoost models presented the lowest prediction error and high efficiency metrics. The most important variable for the prediction models was the daily streamflow on the date of the sediment samples. The average daily rainfall from rain gauge stations was also important for sediment modeling. The use of machine learning models can help sediment monitoring and serve as a great tool for understanding the dynamics of sediment yield in the Doce River Basin over time.

Keywords: Sentinel 2 and Landsat 8. Sediments modeling. Supervised learning models. Fundão tailings dam.

SUMÁRIO

INTRODUÇÃO GERAL

O transporte de sedimentos em bacias hidrográficas tropicais representa cerca de 50% do fluxo de material sólido presente nas águas das áreas continentais que atinge os oceanos (VILLAR et al., 2013). Entender o processo de produção, transporte e deposição de sedimentos é de grande interesse em diversas áreas, pois a concentração de sedimentos, tanto de leito como em suspensão, está diretamente relacionada a problemas com: qualidade da água, perda do volume útil de água nos reservatórios e assoreamento de rios, prejuízos no transporte fluvial, qualidade do ambiente aquático e mal funcionamento e redução da vida útil das usinas hidroelétricas (PETERSON et al., 2018; AL-MUKHTAR, 2019; FROMANT et al., 2021).

Os sedimentos também são agentes fixadores para outros poluentes. Produtos químicos podem ser assimilados nas partículas de sedimentos, servindo assim como um potencializador de problemas decorrentes do uso de produtos agrícolas, microrganismos patogênicos e outros resíduos poluentes (CARVALHO et al., 2000; AFAN et al., 2016).

A compreensão da sedimentologia fluvial é bastante complexa, pois envolve vários fatores de ordem física, meteorológica e antropogênica. Diferentemente das medições de vazão fluvial, o monitoramento da concentração de sedimentos em suspensão ou de leito é bem mais trabalhoso e oneroso, demandando campanhas de campo para coleta de amostras e análises de laboratório. Isso dificulta a obtenção de séries históricas continuas, as quais frequentemente são esporádicas e para um curto período de coleta de dados (MALIK; KUMAR; PIRI, 2017; AL-MUKHTAR; AL-YASEEN, 2019).

Com a crescente demanda por informações relacionadas com a qualidade e conservação dos recursos hídricos, há necessidade de intensificar e ampliar as redes de monitoramento em bacias hidrográficas. Porém, o alto custo de instalação, manutenção e coleta de dados tem desencorajado a instalação de novas estações de monitoramento (JAHANDDIDEH-TEHRANI; BOZORG-HADDAD; DALIAKOPOULOS, 2021). É estimado que 75% dos rios do mundo não possui monitoramento contínuo da qualidade de água, e não há perspectivas para a melhoria deste cenário em um futuro próximo (VILLAR et al., 2012). Observa-se que menos de 10% dos rios que desaguam nos oceanos apresentam algum monitoramento de sedimentos (COHEN et al., 2013).

Neste contexto, o monitoramento da concentração superficial de sedimentos (CSS) por meio de sensoriamento remoto orbital tem sido de grande auxílio para contornar este problema, pois viabiliza o monitoramento contínuo e espacial. Os estudos da medição do fluxo de sedimentos em águas continentais realizados no mundo iniciaram-se entre as décadas de 1980

e 1990 (KIRK, 1981; MERTES; SMITH; ADAMS, 1993), inclusive no Brasil (NOVO; HANSOM; P. J. CURRAN, 1989), utilizando principalmente o satélite *Thematic Mapper* (TM)/ Landsat 5. Esses estudos demonstraram a possibilidade de relacionar o conteúdo de sedimentos e a cor das águas continentais detectada pelos satélites, e comprovaram que as propriedades ópticas da água são influenciadas pelo conteúdo e tipo de sedimento.

Diversos estudos objetivaram verificar a sensibilidade da refletância medida por meio de sensoriamento remoto orbital para a estimativa da CSS, tanto em águas oceânicas como continentais (MARTINEZ et al., 2009; FILIZOLA et al., 2011; VILLAR et al., 2012, 2013; PINTO et al., 2014; MARINHO et al., 2018; SANTOS et al., 2018). Grande parte destes trabalhos observaram forte correlação positiva entre a CSS e a refletância medida pelos satélites, principalmente na faixa espectral de 700 a 800 nm. Isso porque a absorção da luz pelas moléculas de água ocorre de forma seletiva à determinado comprimento de onda. Observa-se que esta absorção é muito baixa nas regiões do azul e do verde, passando a ser significativa a partir de 550 nm. Os valores mais altos de absorção ocorrem no final da região do vermelho e início da região no infravermelho próximo (BARBOSA; NOVO; MARTINS, 2019).

Para o contexto de grandes rios o sensor *Moderate Resolution Imaging Spectroradiometer* (MODIS) tem boa representatividade, com escala de pixel de 250 m e imagens praticamente diárias (NASA, 2018a). No entanto, a sua resolução espacial demonstra-se um fator limitante para sua aplicabilidade em rios com larguras menores, devido à dificuldade de obtenção de pixels que contenham apenas a reflectância da água.

Neste cenário, sensores que apresentam resolução espacial melhor vêm ganhando aplicabilidade para o monitoramento dos fluxos de sedimentos, com destaque para a constelação do sensor *MultiSpectral Instrument* (MSI)/Sentinel 2, disponível a partir de 2015, com resolução espacial de 10 m para as bandas do visível e resolução temporal de 5 dias (ESA, 2018). O sensor *Operational Land Imager* (OLI)/Landsat 8 também tem sido empregado para a quantificação da CSS. Neste sensor a resolução radiométrica passou de 8 bits para 12 bits, o que permite melhor capacidade de detectar as diferenças na energia refletida (NASA, 2018b).

A predição da CSS é fundamental para o gerenciamento e a sustentabilidade dos recursos hídricos. No entanto, a inter-relação de fatores físicos e climáticos fazem da dinâmica dos sedimentos em rios não somente difícil de ser entendida, mas também de ser simulada (BHARTI et al., 2017). Isso ocorre, principalmente, por causa de sua alta variabilidade e do comportamento não estacionário. Desta forma, a predição da concentração de sedimentos requer modelos que possam ter um bom desempenho mesmo com dados faltantes e relação não linear com as variáveis explicativas (MALIK; KUMAR; PIRI, 2017). Devido a essa

complexidade, as técnicas utilizadas para modelar esse fenômeno tem demonstrado pouca capacidade preditiva (MUSTAFA, 2016), e não há uma aceitação universal dos modelos propostos (LAFDANI; NIA; AHMADI, 2013; MALIK; KUMAR; PIRI, 2017).

Os modelos baseados em aprendizagem de máquina trouxeram novas perspectivas para a modelagem da CSS (AFAN et al., 2016). Esses modelos têm apresentado boa capacidade para trabalhar a alta complexidade e não estacionariedade dos dados de sedimentos (NOURANI et al., 2014), permitindo relacionar fatores intrínsecos da área, como relevo, tipos de solos, geologia, mudanças no uso da terra, bem como informações hidrológicas para a quantificação da CSS (RESTREPO; ESCOBAR, 2018).

A região de estudo compreende a bacia hidrográfica do rio Doce, situada na região Sudeste do Brasil. Após o rompimento da barragem de rejeitos de Fundão, ocorrido em 05 de novembro de 2015, cerca de 34 milhões de m³ de rejeitos de mineração foram lançados no meio ambiente. Esse montante, em grande parte, atingiu 663 km de rios e córregos na bacia hidrográfica do rio Doce, nos estados de Minas Gerais e Espírito Santo (MMA 2016). Essa área já apresenta histórico de degradação ambiental em decorrência das diversas atividades antrópicas desenvolvidas e, após o rompimento da barragem de Fundão, esses problemas foram potencializados.

Neste sentido, no primeiro capítulo deste trabalho intitulado "Modelagem da concentração superficial de sedimentos na bacia hidrográfica do rio Doce utilizando sensoriamento remoto orbital" objetivou-se modelar a concentração superficial de sedimentos ao longo da calha do rio Doce a partir dos sensores da constelação do satélite *MultiSpectral Instrument* (MSI)/Sentinel 2 e *Operational Land Imager* (OLI)/Landsat 8.

O segundo capítulo intitulado "Modelagem da concentração superficial de sedimentos na bacia hidrográfica do rio Doce com base em aprendizado de máquina", teve por objetivo a predição da CSS ao longo da calha do rio Doce por meio dos algoritmos de aprendizado de máquina *Random Forest* (RF), *Cubist, Support Vector Machines* (SVMs), *Extreme Gradient Boosting Machine* (XGBoost) e regressão *Least Absolute Shrinkage and Selection Operator* (LASSO). Em que se utilizou como variáveis preditoras informações de declividade, pedologia, uso e cobertura da terra, precipitação, vazão fluvial, velocidade fluvial, evapotranspiração real, escoamento superficial, umidade do solo, temperatura e NDVI. Essas variáveis são mais simples de serem obtidas, portanto de grande auxílio no monitoramento dos sedimentos na bacia hidrográfica do rio Doce.

REFERÊNCIAS

AFAN, H. A. et al. Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction. **Journal of Hydrology**, v. 541, p. 902–913, 1 out. 2016.

AL-MUKHTAR, M. Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. **Environmental Monitoring and Assessment**, v. 191, n. 11, p. 673, 25 nov. 2019.

AL-MUKHTAR, M.; AL-YASEEN, F. Modeling Water Quality Parameters Using Data-Driven Models, a Case Study Abu-Ziriq Marsh in South of Iraq. **Hydrology**, v. 6, n. 1, p. 24, 17 mar. 2019.

BARBOSA, C.; NOVO, E.; MARTINS, V. **Introdução ao Sensoriamento Remoto de sistemas aquáticos**. 1. ed. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 161p. 2019., 2019.

BHARTI, B. et al. Modelling of runoff and sediment yield using ANN , LS-SVR , REPTree and M5 models. **Hydrology Research**, p. 1489–1507, 2017.

CARVALHO, N. de O. et al. **Guia de práticas sedimentométricas**. Brasília: ANEEL, 2000.

CBH-DOCE. **A bacia hidrográfica do Rio Doce**. Disponível em: <http://www.cbhdoce.org.br/institucional/a-bacia>. Acesso em: 5 out. 2018.

COHEN, S. et al. WBMsed, a distributed global-scale riverine sediment flux model: Model description and validation. **Computers & Geosciences**, v. 53, p. 80–93, 1 abr. 2013.

ELESBON, A. A. A. et al. Multivariate statistical analysis to support the minimum streamflow regionalization. **Engenharia Agrícola**, v. 35, n. 5, p. 838–851, out. 2015.

ESA. **Spatial Resolutions Sentinel-2 MSI**. Disponível em: <https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/spatial>. Acesso em: 7 out. 2018.

FILIZOLA, N. et al. The Significance of Suspended Sediment Transport Determination on the Amazonian Hydrological Scenario. **Sediment Transport in Aquatic Environments**, p. 45–64, 2011.

FROMANT, G. et al. Suspended sediment concentration field quantified from a calibrated MultiBeam EchoSounder. **Applied Acoustics**, v. 180, p. 108107, 1 set. 2021.

JAHANDDIDEH-TEHRANI, M.; BOZORG-HADDAD, O.; DALIAKOPOULOS, I. N. The Role of Water Information and Data Bases in Water Resources Management. In: BOZORG-HADDAD, O. (Ed.). **Essential Tools for Water Resources Analysis, Planning, and Management**. Singapore: Springer Singapore, 2021. p. 59–83.

KIRK, J. Monte Carlo study of the nature of the underwater light field in, and the relationships between optical properties of, turbid yellow waters. **Australian Journal of Marine and Freshwater Research**, v. 32, n. 4, p. 517–532, 1981.

LAFDANI, E. K.; NIA, A. M.; AHMADI, A. Daily suspended sediment load prediction using artificial neural networks and support vector machines. **Journal of Hydrology**, v. 478, p. 50–62, 2013.

MALIK, A.; KUMAR, A.; PIRI, J. Daily suspended sediment concentration simulation using hydrological data of Pranhita River Basin, India. **Computers and Electronics in Agriculture**, v. 138, p. 20–28, 2017.

MARINHO, T. et al. Suspended Sediment Variability at the Solimões and Negro Confluence between May 2013 and February 2014. **Geosciences**, v. 8, n. 7, p. 265, 19 jul. 2018.

MARTINEZ, J. M. et al. Increase in suspended sediment discharge of the Amazon River assessed by monitoring network and satellite data. **Catena**, v. 79, n. 3, p. 257–264, 2009.

MERTES, L. A. K.; SMITH, M. O.; ADAMS, J. B. Estimating Suspended Sediment Concentrations in Surface Waters of the Amazon River Wetlands from Landsat Images. **Remote Sens. Environ.**, v. 43, n. 281–301, 1993.

MUSTAFA, M. R. Modeling daily suspended sediments of a hyper-concentrated river in Malaysia. **ARPN Journal of Engineering and Applied Sciences**, v. 11, n. 4, p. 2141–2145, 2016.

NASA. **MODIS Land Science Team**. Disponível em: <https://modis-land.gsfc.nasa.gov/>. Acesso em: 7 out. 2018a.

NASA. **History Landsat Science**. Disponível em: <https://landsat.gsfc.nasa.gov/about/history/>. Acesso em: 29 jan. 2018b.

NOURANI, V. et al. Applications of hybrid wavelet – Artificial Intelligence models in hydrology : A review. **Journal of Hydrology**, v. 514, p. 358–377, 2014.

NOVO, E. M. M.; HANSOM, J. D.; P. J. CURRAN. The effect of sediment type on the relationship between reflectance and suspended sediment concentration. **International Journal of Remote Sensing**, v. 10, n. 7, p. 1283–1289, 1989.

PETERSON, K. T. et al. Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine. **Remote Sensing**, v. 10, n. 10, p. 1–17, 2018.

PINTO, C. E. T. et al. Uso de imagens MODIS no monitoramento do fluxo de sedimentos no reservatório de Três Marias. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 18, n. 5, p. 507–516, 2014.

RESTREPO, J. D.; ESCOBAR, H. A. Sediment load trends in the Magdalena River basin (1980–2010): Anthropogenic and climate-induced causes. **Geomorphology**, v. 302, p. 76–91, 2018.

SANTOS, A. L. M. R. et al. Purus River suspended sediment variability and contributions to the Amazon River from satellite data (2000–2015). **Comptes Rendus Geoscience**, v. 350, n. 1–2, p. 13–19, 1 jan. 2018.

VILLAR, R. E. et al. The integration of field measurements and satellite observations to determine river solid loads in poorly monitored basins. **Journal of Hydrology**, v. 444–445, p. 221–228, 2012.

VILLAR, R. E. et al. A study of sediment transport in the Madeira River, Brazil, using MODIS remote-sensing images. **Journal of South American Earth Sciences**, v. 44, p. 45–54, 2013.

CAPÍTULO 1:

Modelagem da concentração superficial de sedimentos na bacia hidrográfica do rio Doce utilizando sensoriamento remoto orbital

RESUMO: Entender a dinâmica da produção e transporte de sedimentos é imprescindível para o planejamento e gestão dos recursos hídricos de bacias hidrográficas. O monitoramento da concentração superficial de sedimentos (CSS) por meio de sensoriamento remoto orbital contribui para melhorar a compreensão dessa dinâmica, pois possibilita a obtenção de informação a respeito da CSS ao longo da calha do rio e um monitoramento contínuo e espacial. Neste contexto, o objetivo deste trabalho foi modelar a concentração superficial de sedimentos na calha principal do rio Doce a partir dos sensores da constelação do satélite *MultiSpectral Instrument* (MSI)/Sentinel 2 e *Operational Land Imager* (OLI)/Landsat 8. Foram utilizados dados observados médios da CSS da seção transversal medidos em sete estações sedimentométricas da Agência Nacional de Águas e Saneamento Básico (ANA) e em outras sete estações sedimentométricas da Fundação Renova, implantadas em decorrência do rompimento da barragem de rejeitos de Fundão. Como variáveis explicativas foram utilizados os dados de refletância dos sensores orbitais MSI/Sentinel 2 e OLI/Landsat 8 e índices espectrais relacionados com o monitoramento dos recursos hídricos. Foram empregados modelos de regressão linear simples e múltipla para a predição da CSS. Dentre os modelos lineares que utilizam múltiplas variáveis, foram adotados o modelo de regressão linear múltipla (RLM), regressão *Least Absolute Selection Shrinkage Operator* (LASSO) e regressão *Elastic Net*. A banda do infravermelho próximo apresentou forte relação linear com a CSS, tanto utilizando o satélite MSI/Sentinel 2 quanto o OLI/Landsat 8. A regressão linear múltipla e as regressões LASSO e *Elastic Net* apresentaram bom desempenho para a predição da CSS. Os mapas de fluxos de sedimentos indicam redução da CSS na calha do rio Doce em anos mais recentes, o que pode ser indicativo de que parte do material oriundo do rompimento da barragem de rejeitos de Fundão pode ter sido transportado pelos processos de ressuspensão e transporte de sedimentos.

Palavras-chaves: Satélites Sentinel 2 e Landsat 8, rompimento da barragem de rejeitos de Fundão, modelagem hidrossedimentológica

1.1. INTRODUÇÃO

A dinâmica da produção de sedimentos nas bacias hidrográficas tropicais representa cerca de 50% do fluxo de material sólido fluvial que atinge os oceanos, constituindo importante fonte de nutrientes para a vida aquática dos rios continentais e áreas costeiras (LI et al., 2018; VILLAR et al., 2013). Apesar de ser um processo natural, a intensificação das atividades antropogênicas nas bacias hidrográficas, atrelado aos fatores intrínsecos da área, como relevo, tipos de solos e geologia, têm resultado em um aumento expressivo do aporte de sedimentos (LI et al., 2020; RESTREPO; ESCOBAR, 2018).

A concentração de sedimentos está ligada a problemas relacionados com: qualidade da água, perda do volume útil de água nos reservatórios e assoreamento de rios, prejuízos no transporte fluvial e mal funcionamento e redução da vida útil das estruturas das usinas hidroelétricas (PETERSON et al., 2018; AL-MUKHTAR, 2019; FROMANT et al., 2021). Além disso, os sedimentos em suspensão podem agir como um poluente físico, tanto pelo aumento da turbidez, quanto pela capacidade de transporte de poluentes químicos, principalmente nas partículas mais finas (AFAN et al., 2016).

O monitoramento dos sedimentos é imperativo para a mitigação destes problemas e para a preservação dos recursos hídricos. No entanto, as medições em situ são onerosas, pois demandam campanhas de campo e análises laboratoriais e, desta forma, são frequentemente esporádicas, pontuais e para um curto período de tempo dificultando, portanto, a obtenção de séries históricas contínuas (MALIK; KUMAR; PIRI, 2017; AL-MUKHTAR; AL-YASEEN, 2019). Por isso, a representação espacial e temporal da dinâmica dos sedimentos a partir dessas medidas é bastante limitada (KUMAR et al., 2016). No Brasil a rede primária de estações sedimentométricas apresenta densidade de uma estação para cada 17.000 km², com medições trimestrais e muitas séries de dados com expressivas falhas (NAVRATIL et al., 2011).

A utilização de métodos para a estimativa indireta da concentração de sedimentos tem auxiliado na ampliação da disponibilidade de dados. O monitoramento da concentração superficial de sedimentos (CSS) por meio de sensoriamento remoto orbital é uma das técnicas que tem sido muito empregada, pois viabiliza o monitoramento contínuo e espacial. A CSS está relacionada, principalmente, com as bandas da região do espectro visível e infravermelho próximo, isso devido ao elevado coeficiente de absorção da água em comprimentos de onda acima de 750 nm (BARBOSA; NOVO; MARTINS, 2019).

Diversos trabalhos têm obtido resultados promissores no monitoramento da CSS em grandes rios, utilizando principalmente o sensor *Moderate Resolution Imaging*

Spectoradiometer (MODIS) a bordo dos satélites AQUA e TERRA do programa *Earth Observation System* (EOS) (MARTINEZ et al., 2009; PARK; LATRUBESSE, 2014, 2015; ESPINOZA-VILLAR et al., 2018; MARINHO et al., 2018; GALLAY et al., 2019; ZAHIRI; MOLLAEE; ANSARI, 2020).

Esses trabalhos têm demostrado boa concordância entre a reflectância medida pelo sensor com a CSS observada. Apesar do sensor MODIS ter boa representatividade, com escala de pixel de 250 m e imagens praticamente diárias (NASA, 2018a), a sua resolução espacial é um fator limitante para sua aplicabilidade em rios com larguras menores. Isso devido à dificuldade de obtenção de pixels que contenham apenas a reflectância da água (SABERIOON et al., 2020).

Neste cenário, sensores que apresentam resolução espacial maior vem sendo aplicados para o monitoramento dos fluxos de sedimentos, com destaque para a constelação do sensor *MultiSpectral Instrument* (MSI)/Sentinel 2, disponível a partir de 2015, com resolução espacial de 10 m para as bandas do visível e infravermelho próximo, e resolução temporal de cinco dias, após o lançamento do satélite MSI/Sentinel 2B em 2017 (ESA, 2018). Esse satélite apresenta quatro bandas adicionais entre a região do espectro visível e infravermelho próximo (VNIRs), com resolução de 20 m, as quais também tem apresentado boa correlação entre a refletância e a CSS (SABERIOON et al., 2020).

A série de satélites Landsat também tem sido frequentemente utilizada para o monitoramento dos sedimentos. Apesar de apresentar resolução temporal em torno de 16 dias e resolução espacial de 30 m nas bandas da região do visível e infravermelho próximo, este satélite permite a obtenção de séries históricas mais longas, pois estão em operação desde 1972. O sensor *Operational Land Imager* (OLI)/Landsat 8, apresentou melhorias na resolução radiométrica em relação às versões anteriores, passando de 8 para 12 bits. Isso permite melhor capacidade de detectar as diferenças na energia refletida (NASA, 2018b). Além disso, o lançamento do *Operational Land Imager* (OLI 2)/Landsat 9, ocorrido em setembro de 2021, diminuiu o tempo de revisita da constelação desse satélite para oito dias (NASA, 2021), o que é uma vantagem para o monitoramento dos recursos hídricos. Bons resultados foram observados no monitoramento da CSS utilizando o OLI/Landsat 8, com boa correlação principalmente com as bandas do vermelho e infravermelho próximo (PETERSON et al., 2018; YEPEZ et al., 2018; JALLY; MISHRA; BALABANTARAY, 2021).

A utilização de sensoriamento remoto orbital para o monitoramento dos sedimentos na bacia hidrográfica do rio Doce é fundamental para ampliar as bases de dados existentes e melhorar o monitoramento dos sedimentos ao longo da bacia. A região de estudo apresenta

grande produção de sedimentos, tanto que o valor médio observado da CSS foi 386,3 mg/L na estação sedimentométrica de Colatina (56994500), que apresenta área de drenagem de 75,8 mil km² e se encontra próxima à foz do rio Doce, enquanto que na estação sedimentométrica de Óbidos (17050001), localizada na foz do rio Amazonas, com 4,7 milhões de km² de área de drenagem, a concentração média é de 99,2 mg/L (LIMA et al., 2005). A degradação das pastagens e o desmatamento tem potencializado a produção de sedimentos na bacia, que possui problemas sérios de assoreamento (ECOPLAN-LUME, 2010).

O rompimento da barragem de Fundão, ocorrido em 5 de novembro de 2015, lançou cerca de 34 milhões de m³ de rejeitos de mineração no meio ambiente. Esse montante, em grande parte, atingiu 663 km de rios e córregos na bacia hidrográfica do rio Doce, nos estados de Minas Gerais e Espírito Santo (MMA 2016), potencializando a presença de sedimentos em suspensão na bacia. Neste contexto, o objetivo deste trabalho foi modelar a concentração superficial de sedimentos na bacia hidrográfica do rio Doce a partir dos sensores da constelação do satélite Sentinel 2 e Landsat 8, visando a quantificação temporal e espacial dos sedimentos ao longo da calha principal do rio Doce.

1.2. MATERIAL E MÉTODOS

1.2.1. Região de estudo

A área de estudo é a bacia hidrográfica do rio Doce, situada na região Sudeste do Brasil (Figura 1.1), com foco na calha principal do rio Doce, onde estão instaladas a maior parte das estações sedimentométricas da bacia. A bacia possui área de drenagem de, aproximadamente, 86,715 km², sendo 86% no estado de Minas Gerais e 14% no estado do Espírito Santo. O rio Doce tem extensão de 879 km e suas nascentes estão localizadas no estado de Minas Gerais, nas serras da Mantiqueira e do Espinhaço, e sua foz no oceano Atlântico na localidade de Vila Resende, no município de Linhares, Espírito Santo. Cerca de 98% de seu território está inserido no bioma Mata Atlântica (CBH-DOCE, 2018).

O clima predominante na bacia hidrográfica do rio Doce é caracterizado como tropical úmido, com temperatura média anual de 18 ºC (LYRA; RIGO, 2019). A precipitação média anual varia entre 836 mm e 1664 mm, com semestre chuvoso entre os meses de outubro a março. As vazões máximas ocorrem nos meses de dezembro, janeiro e março, e as vazões mínimas, nos meses de agosto e setembro (ECOPLAN-LUME, 2010). Os principais usos e cobertura da terra são agropecuária (63,5%) e floresta plantada e nativa (32,4%)

(MAPBIOMAS, 2022). Os solos predominantes são o Latossolo Vermelho Amarelo (LVA), Argissolo Vermelho (PVe) e Cambissolos Háplicos (CXbe) (IBGE, 2019).

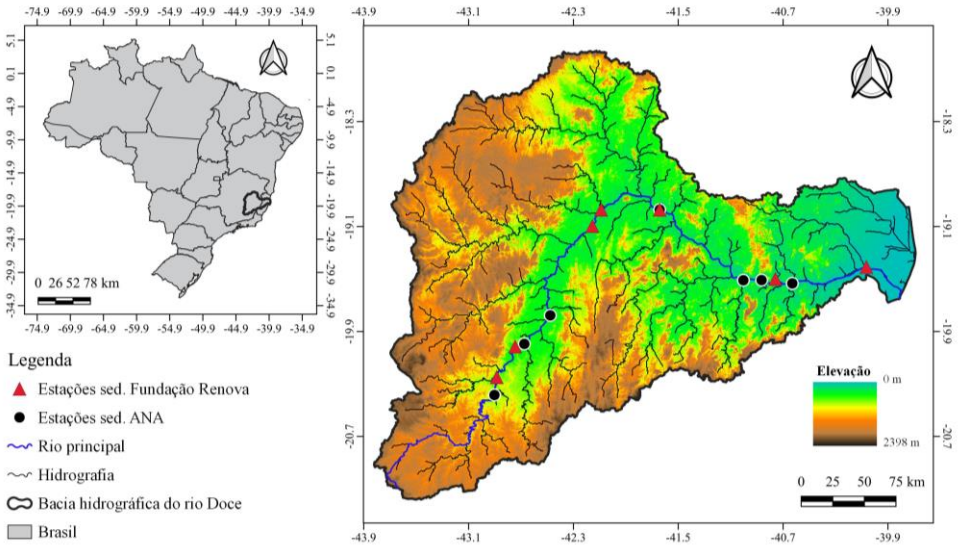


Figura 1.1. Localização da bacia hidrográfica do rio Doce e, Modelo Digital de Elevação Hidrograficamente Condicionado (MDEHC), com destaque para a hidrografia e localização das estações sedimentométricas da Agência Nacional de Águas e Saneamento Básico (ANA) e da Fundação Renova na calha principal do rio Doce.

1.2.2. Sensores orbitais utilizados e obtenção da reflectância da água

Foram utilizadas imagens da constelação MSI/Sentinel 2 com correções atmosféricas (*level* 2A), para evitar a influência de efeitos de espalhamento da atmosfera (luz difusa), como nuvens, aerossóis e gases (HARMEL et al., 2018). A verificação da disponibilidade das imagens para cada período de interesse foi feita na plataforma computacional Google Earth Engine (GEE) (GEE, 2022) e a obtenção das imagens por meio do portal *Copernicus Open Access Hub* (ESA, 2022). Isso porque as correções atmosféricas requerem metadados das imagens que não foram possíveis de serem obtidos por meio da plataforma GEE.

As imagens com correções atmosféricas estão disponíveis a partir de 2019. Para o período anterior foi utilizado o algoritmo Sen2Cor no software SNAP versão 8.0. O Sen2cor, é o mesmo algoritmo utilizado nas imagens já corrigidas (WARREN et al., 2019). Foi feito uma comparação da imagem corrigida pela *European Space Agency* (ESA) e pelo software SNAP versão 8.0, e as métricas de avaliação são apresentadas no Apêndice A. A descrição técnica completa das bandas da constelação do satélite MSI/Sentinel 2 é apresentada no Anexo A.

Foram também utilizadas imagens do satélite OLI/Landsat 8. Esse satélite apresentou evolução em termos de características radiométricas em relação às suas versões anteriores, em decorrência da quantização de 12 bits, permitindo melhor capacidade de detectar as diferenças na energia refletida, o que pode ser de grande auxilio em estudos de sensoriamento remoto da água (GERACE; SCHOTT; NEVINS, 2013). Foram utilizadas as imagens OLI/Landsat 8 *Surface Reflectance* (SR), as quais já possuem correções atmosféricas, obtidas por meio da plataforma GEE. No Anexo B é apresentado as especificações técnicas completas das bandas deste satélite.

Foi obtida a refletância média de cada trecho ao longo da calha principal do rio Doce em que as 14 estações sedimentométricas estão instaladas (Figura 1.1). Para o satélite MSI/Sentinel 2, esse procedimento foi feito utilizando o software QGIS versão 3.16.9 (QGIS DEVELOPMENT TEAM, 2021), em que se recortou as imagens por meio de camadas vetoriais (*shapefiles)* no formato retangular e na posição vertical em relação à margem do rio, com valor mínimo de 12 pixels dependendo da largura do trecho, extraindo-se apenas os pixels inseridos dentro da calha mais profunda do rio, para evitar o efeito da margem. Para o satélite OLI/Landsat 8 foi feito o mesmo procedimento, porém utilizando a plataforma GEE.

Foi avaliada, ainda, a possibilidade da utilização do sensor *Moderate Resolution Imaging Spectoradiometer* (MODIS) a bordo dos satélites AQUA e Terra na bacia hidrográfica do rio Doce. As imagens deste sensor apresentam menor tempo de revisita, possibilitando a obtenção de cenas em até dois dias. No entanto, sua resolução espacial de 250 m nas bandas do visível e infravermelho próximo impossibilitou sua utilização, pois não foi possível a obtenção de imagens com pixels puros de água, mesmo em regiões em que a largura do rio Doce é maior. No Apêndice B é possível observar uma cena do sensor MODIS na região de Linhares, Espírito Santo.

1.2.3. Base de dados da concentração superficial de sedimentos (CSS)

Foram utilizados dados observados da CSS médios da seção transversal medidos em sete estações sedimentométricas administras pela Agência Nacional de Águas e Saneamento Básico (ANA), disponibilizados por meio do portal HidroWeb do Sistema Nacional de Informações sobre Recursos Hídricos (SNIRH) (HIDROWEB, 2022), e também dados de sete estações sedimentométricas da Fundação Renova, responsável pelo monitoramento da qualidade da água para fins da recuperação da bacia hidrográfica do rio Doce após o

rompimento da barragem de Fundão, obtidos através do Portal de Monitoramento do rio Doce (RENOVA, 2022).

A Tabela 1.1 apresenta o código das estações sedimentométricas utilizadas, sua localização geográfica na bacia, descrição, responsável pela operação e largura média do trecho do rio em que a estação está instalada.

Tabela 1.1. Estações sedimentométricas da ANA e Fundação Renova utilizadas

Longitude	Latitude	Código	Nome da estação	Responsável	Largura (m)
-42,903	-20,384	56110005[1]	Ponte Nova Jusante	ANA	52,0
-42,674	-19,994	56425000	Fazenda Cachoerira das Antas	ANA	131,0
-42,476	-19,777	56539000	Cachoeira dos Óculos Montante	ANA	142,0
-41,642	-18,971	56920000	Tumiritinga	ANA	384,0
-41,003	-19,509	56992390	UHE Mascarenhas Montante	ANA	184,0
-40,864	-19,508	56992480	UHE Mascarenhas Jusante	ANA	229,0
-40,630	-19,533	56994500	Colatina	ANA	632,0
-42,885	-20,248	RDO-01[1]	Rio Doce	Renova	50,0
-42,745	-20,014	RDO-03	São Domingos do Prata	Renova	104,0
-42.155	-19,096	RDO-06	Periquito	Renova	338,0
-42,088	-18,971	RDO-07	Governador Valadares	Renova	253,0
-41,642	-18,971	RDO-09	Tumiritinga	Renova	337,0
-40,759	-19,499	RDO-12	Colatina	Renova	224,0
-40,065	-19,408	RDO-15	Linhares	Renova	496,0

[1]não foi possível a utilização dessas estações com OLI/Landsat 8 em decorrência da largura da calha do rio ser insuficiente para obtenção de pixels puros de água.

A frequência das medições nas estações sedimentométricas é de quatro vezes ao ano, o que dificultou a obtenção de imagens dos satélites para a mesma data da coleta. Para contornar esse problema foram consideradas, como critério de seleção, imagens orbitais com até três dias de defasagem em relação a data de coleta das amostras de sedimentos. Esse procedimento foi feito observando a variação da vazão fluvial dentro deste intervalo. Se ocorreram pequenas variações, a concentração de sedimentos tende a não apresentar grandes flutuações (AICH; ZIMMERMANN; ELSENBEER, 2014). Nos Apêndices C e D observa-se a variação da vazão nas estações sedimentométricas da ANA e Fundação Renova no intervalo de três dias antes e após da data da coleta dos sedimentos, nas datas em que foi possível a obtenção das imagens orbitais dos satélites MSI/Sentinel 2 e OLI/Landsat 8, respectivamente.

Nos Apêndices E e F é possível observar as datas das coletas com as respectivas datas de passagem dos satélites MSI/Sentinel 2 e MSI/Landsat 8, respectivamente. Foi possível a utilização de 41 dados observados de sedimentos que atendiam ao critério descrito

anteriormente com o MSI/Sentinel 2, entre 2016 e 2020. No entanto a maior parte das imagens são próximas ao início do segundo semestre de 2017, período em que foi lançado o MSI/Sentinel 2B e diminuiu o tempo de revisita desse satélite. Já para o OLI/Landsat 8 foi possível a utilização de 39 dados, com início no primeiro semestre de 2013 até o segundo semestre de 2020.

1.2.4. Variáveis preditoras utilizadas no ajuste dos modelos de predição da CSS

No ajuste dos modelos foram utilizadas como variáveis preditoras as bandas da região do espectro visível, infravermelho próximo e de ondas curtas dos satélites MSI/Sentinel 2 e OLI/Landsat 8, e também índices espectrais relacionados com sensoriamento remoto da água. Na Tabela 1.2 são apresentadas as bandas espectrais utilizadas como variáveis preditoras.

Tabela 1.2. Especificações das bandas espectrais dos satélites MSI/Sentinel 2 e OLI/Landsat 8 utilizadas como variáveis preditoras.

MSI/ Sentinel 2			
Banda	**Descrição**	**Comprimento de onda (nm)**	**Resolução (m)**
B02	Azul (B)	439 - 535	10
B03	Verde (G)	537 - 582	10
B04	Vermelho (R)	646 - 685	10
B08	Infravermelho Próximo (NIR)	767 - 908	10
B05	Visível e infravermelho próximo (VNIR 1)	694 - 714	20
B06	Visível e infravermelho próximo (VNIR 2)	731 - 749	20
B07	Visível e infravermelho próximo (VNIR 3)	768 - 796	20
B08A	Visível e infravermelho próximo (VNIR 4)	848 - 881	20
B11	Infravermelho de ondas curtas (SWIR 1)	1539 - 168	20
B12	Infravermelho de ondas curtas (SWIR 2)	2072 - 2312	20
OLI/ Landsat 8			
Banda	**Descrição**	**Comprimento de onda (nm)**	**Resolução (m)**
B02	Azul (B)	452 - 512	30 m
B03	Verde (G)	533 - 590	30 m
B04	Vermelho (R)	636 - 673	30 m
B05	Infravermelho próximo (NIR)	851 - 879	30 m
B06	Infravermelho de ondas curtas (SWIR 1)	1567 - 1651	30 m
B07	Infravermelho de ondas curtas (SWIR 2)	2107 - 2294	30 m

Os sedimentos em suspensão estão relacionados principalmente com as bandas da região do espectro visível e infravermelho próximo em decorrência do elevado coeficiente de absorção

da água em comprimentos de onda acima de 750 nm (BARBOSA; NOVO; MARTINS, 2019), e por isso foram selecionadas as bandas em torno desta faixa. As bandas do infravermelho de ondas curtas (SWIR) foram incluídas nas análises por serem utilizadas em alguns índices espectrais da água.

Os índices espectrais utilizados foram os seguintes: *Normalized Difference Water Index* (NDWI), *Modified Normalized*, *Difference Water Index* (MNDWI), *Normalized Difference Turbidity Index* (NDTI), *Water Ratio Index* (WRI), *Automated Water Extraction Index* (AWEI) e *Simple Ratio* (SR), os quais apresentaram bons resultados na modelagem de sedimentos em suspensão segundo SABERIOON et al. (2020). Em geral, os pixels que representam áreas de água exibem valores positivos para os referidos índices, e quando maior este número, menos material em suspensão estão presentes, com exceção do NDTI.

A descrição de cada índice e sua fórmula de cálculo é apresentada na Tabela 1.3. No caso das imagens do satélite MSI/Sentinel 2, o cálculo dos índices espectrais foi feito utilizando o software R versão 4.1.1 (R CORE TEAM, 2021). Já para o satélite OLI/ Landsat 8 o cálculo foi realizado na plataforma GEE.

Tabela 1.3. Índices espectrais utilizados como variáveis preditoras da CSS

Índice espectral	Definição	Referência
MNDWI1	$G - SWIR_1/G + SWIR_1$	Xu (2007)
MNDWI2	$G - SWIR_2/G + SWIR_2$	Xu (2007)
MNDWI3	$G - SWIR_1/G + SWIR_2$	Xu (2007)
MNDWI4	$G - SWIR_2/G + SWIR_1$	Xu (2007)
NDTI	R - G/R + G	Lacaux et al. (2007)
NDWI1	$G - NIR_1/G + NIR_1$	Gao (1996)
NDWI2	$NIR_1 - SWIR_1/NIR_1 + SWIR_1$	Gao (1996)
NDWI3	$NIR_1 - SWIR_2/NIR_1 + SWIR_2$	Gao (1996)
NDWI4	$NIR_1 - SWIR_1/NIR_1 + SWIR_2$	Gao (1996)
NDWI5	$NIR_1 - SWIR_2/NIR_1 + SWIR1$	Gao (1996)
WRI1	$G + R/NIR_1 + SWIR_1$	Mukherjee e Samuel (2016)
WRI2	$G + R/NIR_1 + SWIR_2$	Mukherjee e Samuel (2016)
AWEI1	$4 (G - SWIR_1) - (0.25\ NIR_1 + 2.75\ SWIR_1)$	Feyisa et al. (2014)
AWEI2	$4 (G - SWIR_2) - (0.25\ NIR_1 + 2.75\ SWIR_2)$	Feyisa et al. (2014)
AWEI3	$4 (G - SWIR_1) - (0.25\ NIR_1 + 2.75\ SWIR_2)$	Feyisa et al. (2014)
AWEI4	$4 (G - SWIR_2) - (0.25\ NIR_1 + 2.75\ SWIR_1)$	Feyisa et al. (2014)
SR1	R/NIR	Martinez et al. (2009)
SR2	R/B	Birth e Mcvey (1968)
SR3	G/R	Birth e Mcvey (1968)
SR4[2]	$VNIR_1-((R+VNIR_2)/2)$	Bhangale et al. (2020)
SR5[2]	$VNIR_4/R$	Bhangale et al. (2020)

[2]utilizados apenas com as bandas do satélite MSI/Sentinel 2, pois referem-se as bandas VNIRs

1.2.5. Modelos de regressão e método de validação cruzada utilizados

Foram utilizados modelos de regressão linear simples (RLS) e múltipla para a predição da CSS ao longo da calha do rio Doce. Os modelos foram ajustados utilizando o software R versão 4.1.1, por meio do pacote Caret (KUHN, 2021). O modelo de RLS é descrito pela Equação 1.1,

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \quad (1.1)$$

em que y é a variável resposta, x_1 é o valor da variável preditora, β_0 é a constante de regressão e representa o intercepto, β_1 é o coeficiente angular, e ε é o erro associado a distância entre os valores preditos e os valores observados.

As variáveis preditoras utilizadas nos modelos de RLS foram selecionadas por meio da matriz de correlação de Pearson (r), as quais obtiveram valores de correlação superiores a 0,7 com a CSS. Marinho et al. (2021) observaram valores de r entre os dados de sedimentos em suspensão com a refletância medida pelo satélite MSI/Sentinel 2 variando de 0,67 a 0,92, considerados valores relativamente altos.

Nos Apêndices G e H são apresentadas as matrizes de correlação entre as variáveis preditoras entre si e com a CSS, obtidas a partir das bandas dos satélites MSI/Sentinel 2 e OLI/Landsat 8, respectivamente, e calculadas através do pacote Corrplot (WEI e SIMKO, 2021) do software R versão 4.1.1. A análise da correlação entre as variáveis é importante no caso dos modelos de regressão linear múltipla, com o intuito de evitar problemas de multicolinearidade.

Dentre os modelos lineares que usam múltiplas variáveis, foram utilizados o modelo de regressão linear múltipla (RLM), regressão *Least Absolute Selection Shrinkage Operator* (LASSO) e regressão *Elastic Net*. A RLM é a extensão da regressão linear simples através adição de múltiplas variáveis preditoras (Equação 1.2), com a finalidade de melhorar a capacidade preditiva do modelo, pois a complexidade de determinados fenômenos dificilmente é explicada por apenas uma variável (HAIR et al., 2009).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon \quad (1.2)$$

em que y_i é a variável resposta, x_{i1}, x_{i2}, ..., x_{ip} são as variáveis preditoras e β_0, β_1, ..., β_p são os coeficientes da regressão, ε é o erro associado a distância entre os valores preditos e os valores observados.

Como as bandas dos satélites apresentam forte correlação entre si, é possível a ocorrência de multicolinearidade entre as variáveis explicativas, as quais podem ter pouca influência na redução da soma do quadrado dos erros sendo, portanto, passíveis de serem removidas. Para isso aplicou-se a técnica *Variance Inflation Factor* (VIF), utilizada para diagnosticar a multicolinearidade, utilizando o pacote Car (FOX e WEISBERG, 2019) do software R versão 4.1.1. A VIF foi calculada por meio da Equação 1.3. Valores altos de VIF_i indica que a variável x_i tem dependência linear com pelo menos uma outra variável (ALIN, 2010).

$$VIF_i = \frac{1}{1 - R_i^2} \tag{1.3}$$

em que R_i^2 é o coeficiente de determinação entre as múltiplas variáveis x_i.

Em geral, utiliza-se a VIF com valor máximo de 10, o que indica baixa dependência linear entre as variáveis (ALIN, 2010). Portanto, para a RLM utilizou-se as variáveis explicativas que foram significativas para o modelo ao nível de significância de 95% e que apresentaram valores de VIF inferiores a 10.

A regressão LASSO é usada tanto como modelo de predição como para seleção do conjunto ótimo de variáveis em modelos de aprendizado de máquina. Esse modelo penaliza a inserção de novas varáveis que apresentam pouca capacidade preditiva por meio do processo de regularização, introduzindo viés aos modelos, o que evita problemas de sobreajuste (*overfitting*) e multicolinearidade (JAMES et al., 2021).

A regressão linear ajusta os coeficientes de regressão β_i de modo a minimizar a função de custo, soma residual dos quadrados (SRQ), sendo expressa pela Equação 1.4.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij} \right)^2 \tag{1.4}$$

em que y_i é o valor observado, β_j e β_0 são os coeficientes de regressão, x_{ij} é a variável explicativa.

A regressão LASSO penaliza os coeficientes de regressão (β_j) acrescentado um novo termo na SRQ, conforme a Equação 1.5. Esse termo é conhecido em estatística como penalização do tipo L1, o qual tem o efeito de zerar os coeficientes quando o valor de λ for

suficientemente grande, pois à medida que este aumenta, menor fica a inclinação da reta, sendo particularmente útil quando há um grande número de variáveis passíveis de serem removidas (JAMES et al., 2021).

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{1.5}$$

em que λ é um hiperparâmetro da equação que vária de 0 a $+\infty$, obtido utilizando validação cruzada para identificar o valor de λ que resulta na menor variância.

A regressão *Elastic Net*, por sua vez, é uma combinação da regressão LASSO e da regressão *Ridge*, descrita pela Equação 1.6. (JAMES et al., 2021).

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda_1\sum_{j=1}^{p}\left|\beta_j\right|+\lambda_2\sum_{j=1}^{p}\beta_j^2 \tag{1.6}$$

em que λ_1 e λ_2 são hiperparâmetro das equações de regressão LASSO e Ridge, respectivamente. Estes variam de 0 a $+\infty$, e são obtidos por meio da validação cruzada de forma a minimizar a variância.

A regressão *Ridge* é bastante similar à regressão LASSO, entretanto, a penalização exercida por esta é do tipo L2, ou seja, quanto maior for o λ, o valor do coeficiente da regressão será assintoticamente próximo a zero, não sendo possível a seleção de variáveis por esse método. Porém, este é bastante útil quando as variáveis são importantes para o modelo, mas relativamente correlacionadas entre si, sendo que no caso da regressão LASSO uma delas seria eliminada e na regressão *Ridge* a penalização reduz e aproxima o valor de seus coeficientes (JAMES et al., 2021).

A vantagem da utilização da regressão *Elastic Net* está na combinação desses dois métodos, permitindo eliminar variáveis com pouca capacidade preditiva (penalização L1) e manter aquelas que são importantes para o modelo, porém com certo grau de correlação entre si, por meio da penalização L2, de modo a evitar problemas com multicolinearidade (DE MOL; DE VITO; ROSASCO, 2009).

O treinamento e teste dos modelos foram feitos utilizando o método de validação cruzada *leave-one-out* (LOOCV), implementado utilizando o software R versão 4.1.1, por meio do pacote Caret. Esse método é frequentemente usado em estudos hidrológicos em decorrência

da disponibilidade limitada de dados hidroclimáticos (HADDAD et al., 2013). O LOOCV é um caso especial de validação cruzada *k-fold*, o qual também envolve a divisão da base de dados em duas partes para treinamento e teste, no entanto, ao invés de criar dois subconjuntos, apenas uma observação (x_1, y_1) é usada para o teste e o restante das observações {(x_2, y_2),..., (x_n, y_n)} são empregadas no treinamento. O modelo de predição utilizado é ajustado em n -1 observações, e uma predição $\hat{y}_1$ é feita para a observação x_1 que não entrou no conjunto de treinamento (JAMES et al., 2021). Esse procedimento é repetido até que todo o conjunto de dados tenha sido utilizado para treinamento e teste.

A vantagem desse método consiste na redução do viés em função da repetição do ajuste do modelo em n-1 vezes. O modelo preditivo gerado é mais estável, pois a divisão dos subconjuntos de treinamento não é feita de forma aleatória, como é o caso dos demais tipos de *k-fold,* em que os resultados podem ser diferentes dependendo do conjunto de dados selecionado. Entretanto, para um grande conjunto de dados o método LOOCV demanda computadores com grande capacidade de processamento (JAMES et al., 2021).

1.2.6. Métricas para avaliação dos modelos de predição da CSS

As métricas utilizadas para avaliação dos modelos em relação a concordância dos dados preditos com os observados da CSS foram: Erro Médio Absoluto (MAE), Raiz do Erro Médio Quadrático (RMSE), Porcentagem do Viés (PBIAS), coeficiente de Nash–Sutcliffe (NSE), índice de concordância de Willmot (d), coeficiente de determinação (R^2), coeficiente de Kling-Gupta (KGE) e índice de desempenho (c). As métricas de avaliação dos modelos foram obtidas utilizando o software R versão 4.1.1, por meio do pacote hydroGOF (ZAMBRANO-BIGIARINI, 2020).

O Erro médio absoluto (MAE) foi obtido por meio da Equação 1.7. Essa métrica serve como indicativo da presença de outlier nos dados utilizados, em situações em que o RMSE for superior ao MAE. Quanto menor o valor obtido na performance do modelo para essa métrica, mais acurado. O ideal é a obtenção de valores próximo a zero (LEGATES; MCCABE, 1999).

$$\mathrm{MAE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{n}} \left| \mathrm{O_i} - \mathrm{P_i} \right| \tag{1.7}$$

em N é o número de elementos da amostra, P_i é o valor dos dados preditos, O_i é o valor dos dados observados.

A RMSE é a medida mais comumente utilizada para aferir a qualidade do ajuste dos modelos, caracterizada por ser uma medida análoga ao desvio padrão e valores do erro nas mesmas dimensões da variável analisada (HALLAK; PEREIRA FILHO, 2011), obtida pela Equação 1.8.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - O_i\right)^2} \tag{1.8}$$

em n é o número de elementos da amostra, P_i é o valor dos dados estimados, O_i é o valor dos dados observados.

A RMSE é uma medida da magnitude média dos erros estimados, em que seu valor é sempre positivo, sendo que quanto mais próximo a zero, melhor é a qualidade dos valores estimados. A desvantagem deste índice se configura no fato de que bastam alguns valores discrepantes para que ocorra um aumento significativo em sua magnitude (SANTOS et al., 2014).

A Pbias indica a diferença entre os valores preditos e observados. Valores baixos de Pbias indicam boa performance do modelo, sendo que zero é o valor ideal. Valores positivos indicam superestimativas, enquanto que valores negativos subestimativas dos modelos (LI et al., 2009). A Equação 1.9 demostra a fórmula de obtenção dessa métrica.

$$\mathrm{Pbias} = \left[\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)100}{\sum_{i=1}^{n} O_i}\right] \tag{1.9}$$

em P_i é o valor dos dados estimados e O_i é o valor dos dados observados.

O NSE é um índice de eficiência adimensional que variam de -∞ a 1, em que NSE igual a 1 indica ajuste perfeito dos dados, enquanto que valores menores que 0, sugerem que a média dos dados observados é melhor que o modelo ajustado (RITTER; MUÑOZ-CARPENA, 2013). Este pode ser obtido com base na Equação 1.10.

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(O_i - P_i\right)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}_i\right)^2} \tag{1.10}$$

em que n é o número de elementos da amostra, P_i é o valor dos dados estimados, O_i é o valor dos dados observados, $\bar{O}_i$ é a média dos valores observados.

O NSE é amplamente utilizado em recursos hídricos devido a sua flexibilidade para trabalhar com diversos modelos matemáticos. Este índice tem maior sensibilidade ao viés nas predições dos modelos, apresentando resultados mais realistas (ALTHOFF; RODRIGUES, 2021).

O índice de concordância (d) proposto por Wilmott é empregado para identificar o grau de concordância entre o valor observado e sua estimativa, e pode ser obtido por meio da Equação 1.11.

$$d = 1 - \left[\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{\sum_{i=1}^{n}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} \right] \tag{1.11}$$

em que n é o número de elementos da amostra, $\bar{O}$ é a média dos valores observados, Oi representa os dados observados e Pi os valores estimados.

Este índice é de simples interpretação, sendo que seus valores variam de 0 a 1, sendo 1 o ajuste perfeito do modelo utilizado. No entanto, este apresenta a desvantagem quanto ao uso das diferenças quadráticas em seu algoritmo, o qual pode resultar em valores altos (bons ajustes) mesmo se a capacidade preditiva do modelo for ruim (PEREIRA et al., 2018).

O R^2 também é utilizado para verificar o ajuste dos modelos em relação a sua capacidade de predizer um fenômeno. Pode ser expresso por meio da Equação 1.12.

$$R^2 = \frac{\sum\left(\hat{P}_i - \bar{O}\right)^2}{\sum\left(O_i - \bar{O}\right)^2} \tag{1.12}$$

em que $\bar{O}$ é a média dos valores observados, O_i representa os dados observados e Pi os valores estimados.

A interpretação dos resultados obtidos do R^2 são similares aos do coeficiente d, ou seja, variando de 0 a 1, os valores ideais são aqueles próximos da unidade. Uma das principais desvantagens do R^2 é a quantificação apenas de dispersão (variação) dos dados se for considerado isoladamente. Um modelo que sub ou superestima sistematicamente a variável observada ao longo do tempo poderá ter valores de R^2 próximos de 1, mesmo se todas as estimativas estiverem erradas, ou seja, não considera o viés (NAGUETTINI; PINTO, 2007).

O KGE envolve a avaliação de três componentes entre os dados preditos e observados, as quais são: a correlação, viés e medidas de variabilidade (GUPTA et al., 2009), obtido por meio da Equação 1.13. Os valores podem variar de -∞ a 1, sendo 1 o valor do ajuste perfeito dos dados pelo modelo.

$$KGE = 1 - ED = 1 - \sqrt{\left[s_r\left(r-1\right)\right]^2 + \left[s_\alpha\left(\alpha-1\right)\right]^2 + \left[s_\beta\left(\beta-1\right)\right]^2} \quad (1.13)$$

em que ED é a distância euclidiana, r é o coeficiente de correlação entre O_i e P_i, α é a razão entre o desvio padrão dos dados preditos com o desvio padrão dos dados observados, β é a razão entre a média dos valores preditos pela média dos valores observados, s_r, s_α e s_β são fatores de escala.

O índice de eficiência (c) foi proposto por Camargo e Sentelhas (1997) e permite a classificação dos modelos ajustados para dar suporte na escolha daquele que apresentou melhor desempenho. Esse índice é o resultado do produto entre o coeficiente de Willmott (d) e o coeficiente de correlação de Pearson (r), em que valores de $c \leq 0,4$ indicam que os modelos são classificados como péssimos; entre 0,41 a 0,5: mal; entre 0,51 a 0,60: sofrível; entre 0,61 a 0,65: mediano; entre 0,66 a 0,75: bom; entre 0,76 a 0,85: muito bom; e $c > 0,85$ que os modelos são classificados como ótimos.

1.2.7. Mapas de fluxos de sedimentos

Os mapas de fluxo de sedimentos foram gerados a partir do modelo que apresentou o menor erro de predição e melhor concordância entre os valores preditos e observados, avaliados por meio das métricas de desempenho. Os mapas foram calculados utilizando a ferramenta *raster calculator* do software QGIS versão 3.16.9.

Foram gerados mapas de fluxos de sedimentos para o período correspondente ao final de 2015 e 2016, dependendo da disponibilidade de imagens, com intuito de avaliar a CSS logo

após o rompimento da barragem de Fundão. Também foram gerados mapas de fluxos de sedimentos para os anos de 2019 e 2020, com objetivando observar se ocorreu diminuição da CSS anos após do acidente.

1.3 RESULTADOS E DISCUSSÃO

Devido à dificuldade de obtenção de imagens dos satélites MSI/Sentinel 2 e OLI/Landsat 8 que correspondessem à mesma data de coleta das amostras da CSS, foi adotado o critério de utilização de imagens com até, no máximo, três dias de defasagem em relação a data da coleta.

Para tanto, avaliou-se a variação da vazão dentro deste intervalo, pois o aumento da CSS está relacionado com o aumento da vazão, tanto pela produção de sedimento nas bacias hidrográficas com os eventos de precipitação (ZHAO et al., 2019), quanto pelo aumento da velocidade fluvial, que faz com que o material mais fino depositado no leito entre em suspensão (BOGGS, 2006).

Verificando-se a variação da vazão nas estações sedimentométricas no período de três dias antes e após a data da coleta de sedimentos em relação aos dias em que foram possíveis a obtenção das imagens do satélite MSI/Sentinel 2, foi observado que apenas as estações 56425000 e RDO-07 apresentaram variações relativamente altas das vazões, com valores de até 50% maiores do que aquele observado na data da coleta de sedimentos (Apêndice C). Para o satélite OLI/Landsat 8, variações acima da média foram verificadas nas estações 56425000, 56992480, RDO-09, RDO-15 (Apêndice D). Entretanto, para ambos os satélites, os dados observados de sedimentos apresentaram boa correlação com a refletância medida e, por isso, foram mantidos na análise.

Além disso, é observado na relação entre a CSS e a vazão um atraso entre o aumento da vazão e subsequente aumento nos sedimentos em suspensão, conhecido como fenômeno da histerese (CAO et al., 2021; HADDADCHI; HICKS, 2021). Portanto, é esperado que não se tenham grandes flutuações nos dados de sedimentos no período considerado na análise.

1.3.1. Variáveis preditoras utilizadas no ajuste dos modelos de predição da CSS

No ajuste dos modelos de predição da CSS utilizando apenas uma variável foram utilizadas aquelas que apresentaram coeficiente de correlação de Pearson (r) maior ou igual a 0,7, o que indica correlação alta (HAIR et al., 2009). Com base na matriz de correlação

(Apêndice G e H), foram verificadas as bandas e índices espectrais que atenderam a esse critério, as quais podem ser observadas na Tabela 1.4.

Tabela 1.4. Bandas e índices espectrais com correlação superior a 0,7 com a CSS, utilizados como variável preditiva nos modelos de regressão linear simples

MSI/ Sentinel 2	Correlação (r)	β_0	β_1	OLI/ Landsat 8	Correlação (r)	β_0	β_1
B04	0,74	-14,85	477,59	B05	0,91	-24,22	1751,49
B08	0,92	-7,17	945,33	NDWI1	-0,78	166,24	-247,9
B05	0,81	-11,83	500,10	-	-	-	-
B06	0,93	-4,89	832,85	-	-	-	-
B07	0,93	-6,96	814,90	-	-	-	-
B08A	0,96	-7,72	1153,16	-	-	-	-
NDWI1	-0,76	118,74	-169,23	-	-	-	-

Na Tabela 1.4 pode-se notar que o satélite MSI/Sentinel 2 apresentou maior quantidade de bandas com correlação alta em relação aos dados de CSS, tanto que foram selecionadas todas as bandas do visível e infravermelho próximo (VNIRs) (Tabela 1.2). O MSI/Sentinel 2 apresenta grande potencial para a modelagem dos sedimentos por disponibilizar essas bandas adicionais na faixa espectral em que a energia não é totalmente absorvida pela água (BARBOSA; NOVO; MARTINS, 2019; SABERIOON et al., 2020).

Dentre os índices espectrais, apenas o NDWI 1 apresentou correlação superior a 0,7 em ambos os satélites analisados. O valor negativo indica correlação inversa com a CSS, o que é esperado, dado que os valores desse índice diminuem com a presença de materiais em suspensão na água (GAO, 1996).

Nos modelos que utilizam mais de uma variável, utilizou-se a função *varImp,* do pacote Caret do software R versão 4.1.1, para verificar as variáveis mais importantes em cada modelo. Empregando-se as variáveis preditoras derivadas dos satélites MSI/Sentinel 2 e OLI/Landsat 8, aquelas que apresentaram maior importância podem ser observadas na Figura 1.2.

Essa figura representa uma pré-seleção das variáveis preditoras. Como há poucos dados observados, procurou-se eliminar as variáveis com menos importância, mesmo que tivessem sido selecionadas, e verificar o impacto dessa remoção nas métricas de avaliação dos modelos. Em modelos de regressão múltipla recomenda-se a proporção mínima de 5 observações para 1 variável, pois o número de graus de liberdade está ligado à capacidade de generalização dos modelos (HAIR et al., 2009).

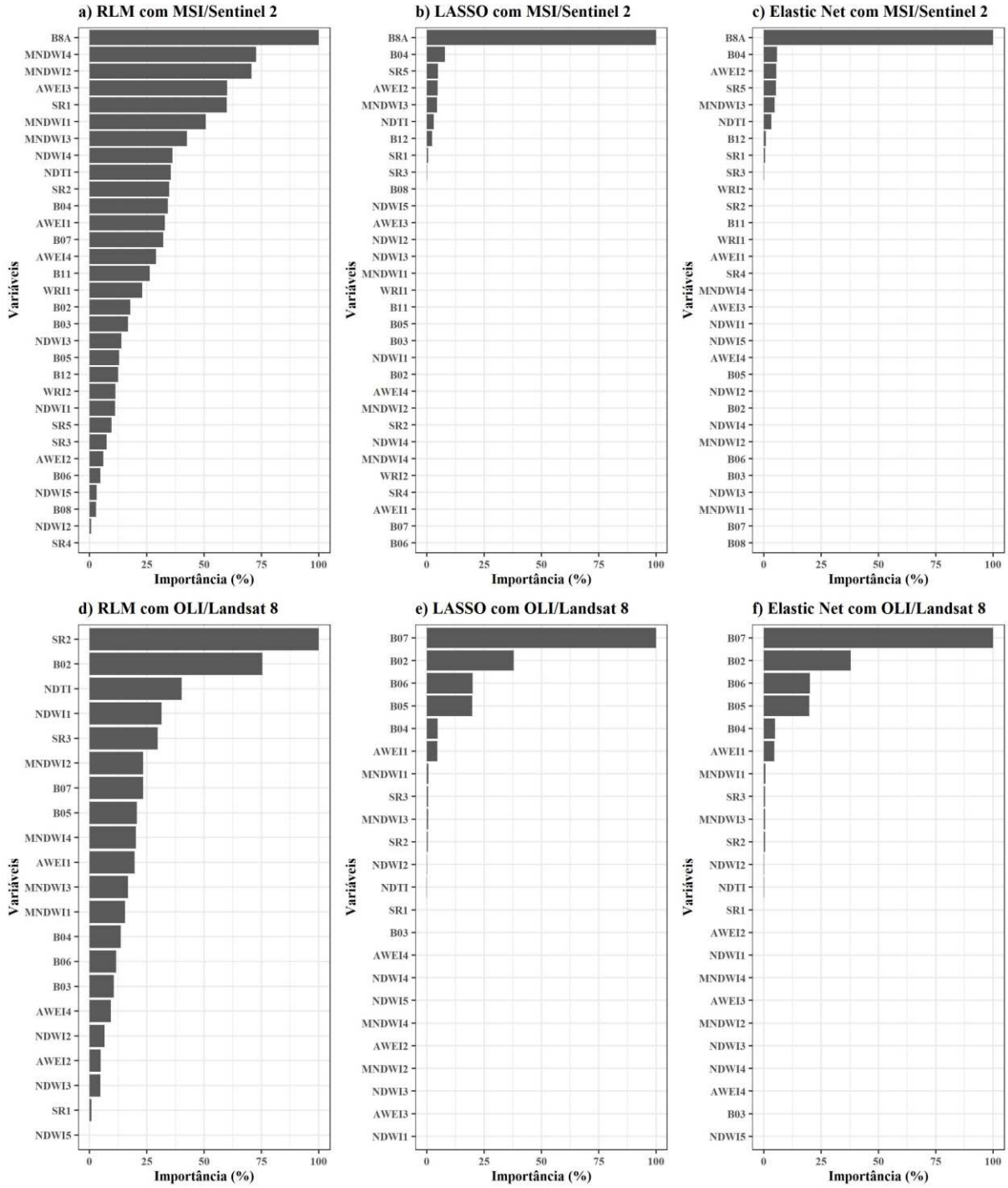


Figura 1.2. Importância das variáveis preditoras derivadas dos satélites MSI/Sentinel 2 e OLI/Landsat 8 para os modelos RLM, LASSO e *Elastic Net*.

Observa-se, na Figura 1.2, que as variáveis selecionadas pela regressão LASSO e regressão *Elastic Net* foram similares, indicando que as variáveis que foram removidas não tinham importância para os modelos, pois a penalização do tipo L1 prevaleceu (JAMES et al.,

2021). As principais variáveis selecionadas por esses modelos de regressão foram bastante similares em ambos os satélites utilizados.

As variáveis selecionadas no modelo de RLM se diferenciaram bastante entre o MSI/Sentinel 2 e o OLI/Landsat 8. No entanto, observa-se que a banda do infravermelho próximo (B08A) e a banda B02 também se destacaram nesse método de regressão para os satélites MSI/Sentinel 2 e OLI/Landsat, respectivamente.

Na Tabela 1.5 pode-se verificar as variáveis que permaneceram no modelo sem alterar sua capacidade preditiva, os valores de VIF nas variáveis utilizadas nos modelos de RLM e os coeficientes de regressão.

Tabela 1.5. Variáveis preditivas finais, os valores de VIF nas variáveis utilizadas nos modelos de RLM e os coeficientes de regressão

MSI/Sentinel 2						
RML			**LASSO**		***Elastic Net***	
Modelo	Coeficientes	VIF	Modelo	Coeficientes	Modelo	Coeficientes
Intercepto	-27,80	-	Intercepto	-30,56	Intercepto	-30,037
B08A	1513,90	5,56	B08A	1300,07	B08A	1190,66
MNDWI4	42,13	2,47	B06	185,79	B06	281,47
SR1	0,81	3,15	B04	-276,86	B04	-291,48
B04	-231,32	4,59	MNDWI3	50,78	MNDWI3	50,48
-	-	-	B12	258,54	B12	300,5
OLI/Landsat 8						
RML			**LASSO**		***Elastic Net***	
Modelo	Coeficientes	VIF	Modelo	Coeficientes	Modelo	Coeficientes
Intercepto	101,48	-	Intercepto	113,47	Intercepto	121,55
B02	-3158,83	6,78	B07	19702,61	B07	16114,97
B05	2973,83	8,87	B02	-6414,27	B02	-6088,24
AWEI1	362,35	10,87	B06	-6251,34	B06	-3695,7
SR2	-57,56	11,58	B05	3313,38	B05	3240,92
-	-	-	B04	-965,39	B04	-894,72
-	-	-	AWEI1	889,89	AWEI1	819,98
-	-	-	MNDWI1	-529,38	MNDWI1	-96,96
-	-	-	SR3	-113,84	SR3	-110,86
-	-	-	MNDWI3	719,01	MNDWI3	279,67
-	-	-	SR2	-82,2	SR2	-79,97

Os modelos que utilizaram as variáveis derivadas do satélite MSI/Sentinel 2 empregaram menos variáveis para a obtenção dos melhores índices na avaliação dos modelos. Apenas nos modelos que empregaram imagens do OLI/Landsat 8 o intercepto foi positivo, ou

seja, apresentou significado físico, em que na situação onde todas as variáveis fossem zero, o valor do intercepto seria o próprio valor da variável estimada (HAIR et al., 2009). As variáveis AWEI1 e SR2 apresentaram valores de VIF superiores a 10, entretanto foram mantidas pois a sua remoção reduziu substancialmente na capacidade preditiva dos modelos.

Grande parte dos estudos que envolvem a modelagem dos sedimentos utilizando sensoriamento remoto orbital apontam que as bandas do vermelho e infravermelho próximo apresentam melhores resultados (MARTINEZ et al., 2009; PARK; LATRUBESSE, 2014, 2015; ESPINOZA-VILLAR et al., 2018; MARINHO et al., 2018; GALLAY et al., 2019; ZAHIRI; MOLLAEE; ANSARI, 2020). Entretanto, as propriedades ópticas da água são influenciadas pelo conteúdo e tipo de sedimentos (MARTINEZ et al., 2009) e, desta forma, diferentes bandas podem ter bom desempenho, dependendo da resposta espectral da composição do material em suspensão presente na água.

No caso deste trabalho, nota-se que a banda do infravermelho próximo (B08A) foi a mais importante para a predição da CSS no satélite MSI/Sentinel 2 em todos os modelos de regressão linear múltipla utilizados, com maior coeficiente de regressão, seguido da banda do vermelho (B04), nos modelos de RLM e LASSO. No estudo de Marinho et al. (2021), utilizando o satélite MSI/Sentinel 2 em rios da região amazônica, os melhores resultados foram obtidos utilizando a banda (B04), com coeficiente de determinação de 0,84. Entretanto, a banda B08A também apresentou métricas aceitáveis.

As bandas do infravermelho de ondas curtas (B11 e B12 para o satélite MSI/Sentinel 2, e B06 e B07 para o OLI/Landsat 8) apresentaram grande importância na modelagem da CSS ao longo da calha do rio Doce, em especial utilizando o satélite OLI/Landsat 8. Não é usual a utilização destas bandas em sensoriamento remoto da água, pois em seu comprimento de onda, a energia incidente é totalmente absorvida pela água. No entanto, em rios em que a turbidez é muito alta, com a CSS superior a 100 mg/L, estas bandas podem ter grande aplicabilidade, uma vez que a banda do infravermelho próximo pode atingir a saturação nestas condições (KNAEPS et al., 2015).

Nos modelos LASSO e *Elastic Net* para o satélite OLI/Landsat 8, destacou-se ainda a banda do azul (B02) como grande importância para a predição da CSS. Em geral essa banda é emprega em estudos do estado trófico de rios e reservatórios, em especial na avalição da presença de clorofila-a. Entretanto, a banda B02 também foi utilizada como variável preditora em modelos de regressão múltiplas para predição da CSS (ANDRZEJ URBANSKI et al., 2016; JAELANI et al., 2016; ZHAO et al., 2020).

1.3.2. Avaliação dos modelos de predição da CSS

Por meio do treinamento e teste dos modelos empregando a validação cruzada LOOCV, obteve-se os modelos de predição da CSS para calha do rio Doce. Na Tabela 1.6 pode-se observar os resultados das métricas de avaliação para cada modelo utilizado (Tabelas 1.4 e 1.5).

Tabela 1.6. Resultados das métricas de avaliação dos modelos utilizando as variáveis derivadas dos satélites MSI/Sentinel 2 e OLI/Landsat 8

MSI/Sentinel 2								
Modelos	**MAE**	**RMSE**	**PBIAS**	**NSE**	**d**	**R²**	**KGE**	**c**
RLS (B04)	23,61	41,60	-1,20	0,39	0,75	0,39	0,53	0,47
RLS (B08)	15,86	26,01	-2,20	0,76	0,93	0,76	0,81	0,81
RLS (B05)	21,59	37,46	-1,90	0,50	0,82	0,51	0,63	0,36
RLS (B06)	14,96	23,25	-2,30	0,81	0,94	0,81	0,83	0,85
RLS (B07)	15,21	24,47	-2,40	0,79	0,93	0,79	0,82	0,83
RLS (B08A)	11,71	18,17	-1,70	0,88	0,97	0,88	0,88	0,91
RLS (NDWI1)	25,00	40,47	-2,50	0,42	0,77	0,42	0,55	0,50
RLM	11,51	14,24	-1,00	0,93	0,98	0,93	0,93	0,94
LASSO	11,90	15,17	-2,90	0,92	0,98	0,92	0,91	0,94
Elastic Net	11,91	15,15	-2,80	0,92	0,98	0,92	0,91	0,94
OLI/Landsat 8								
Modelos	**MAE**	**RMSE**	**PBIAS**	**NSE**	**d**	**R²**	**KGE**	**c**
RLS (B05)	28,75	47,16	-2,20	0,75	0,92	0,75	0,81	0,80
RLS (NDWI1)	45,94	67,81	-3,30	0,48	0,80	0,48	0,60	0,56
RLM	26,96	41,95	-1,70	0,80	0,94	0,80	0,86	0,84
LASSO	21,96	29,01	-3,00	0,90	0,97	0,91	0,91	0,92
Elastic Net	21,41	28,66	-2,50	0,91	0,97	0,91	0,91	0,92

A Tabela 1.6 demonstra que para o satélite MSI/Sentinel 2, o modelo RLM apresentou desempenho superior aos demais utilizados, tanto em relação aos índices que indicam a eficiência dos modelos como nas métricas de erro. Entretanto, a seleção das variáveis finais do modelo é mais trabalhosa, enquanto a regressão LASSO e *Elastic Ne* reduzem bastante o número de variáveis. Em um contexto com muitas variáveis preditoras, é importante considerar a utilização desses métodos (JAMES et al., 2021).

A banda do infravermelho próximo com resolução espacial de 20 m (B08A) como variável preditora apresentou bom desempenho em comparação com demais modelos que utilizaram apenas uma variável, até mesmo em relação a banda do infravermelho próximo com 10 m de resolução (B08). Isso pode ser evidenciado pelo o índice de eficiência (c), que para o

modelo de RLS utilizando esta banda é classificado como ótimo ($c \geq 86$), assim como os modelos de regressão que utilizam múltiplas variáveis. Isso demonstra o potencial da utilização do satélite MSI/Sentinel 2 na modelagem de sedimentos a partir da inclusão das bandas do visível e infravermelho próximo (VNIRs), conforme também indicado por SABERIOON et al. (2020).

A predição da CSS utilizando variáveis derivadas do satélite OLI/Landsat 8 apresentou limitações relacionadas à resolução espacial. As estações com código RD01 e 56110005, localizadas na região próxima ao local de desague dos rejeitos da barragem de Fundão, não foram utilizadas em decorrência da menor largura do rio, que impossibilitou a obtenção de pixels apenas inseridos na calha do rio. Apesar da resolução espacial desse satélite ser de 30 m, é preciso considerar o efeito da margem que, em geral, é menos profunda e pode ter influência na refletância (BARBOSA; NOVO; MARTINS, 2019).

Bons resultados também foram obtidos utilizando as variáveis preditoras derivadas do satélite OLI/Landsat 8. O modelo de regressão *Elastic Net* apresentou capacidade de predição melhor que os demais modelos utilizados, no entanto, ligeiramente superior à regressão LASSO. Entretanto, foi necessário a utilização de uma quantidade maior de variáveis preditoras em comparação com o ajuste obtido com as variáveis do satélite MSI/Sentinel 2. A vantagem da utilização do OLI/Landsat 8 é, principalmente, a possibilidade de obtenção de séries de dados mais longas (PETERSON et al., 2018), como foi o caso desse estudo, em que obteve-se imagens a partir de 2013.

Nas Figuras 1.3 e 1.4 é possível observar a relação entre os valores preditos e observados obtidos através das variáveis preditoras e diferentes modelos, utilizando o satélite MSI/Sentinel 2 e OLI/Landsat 8, respectivamente.

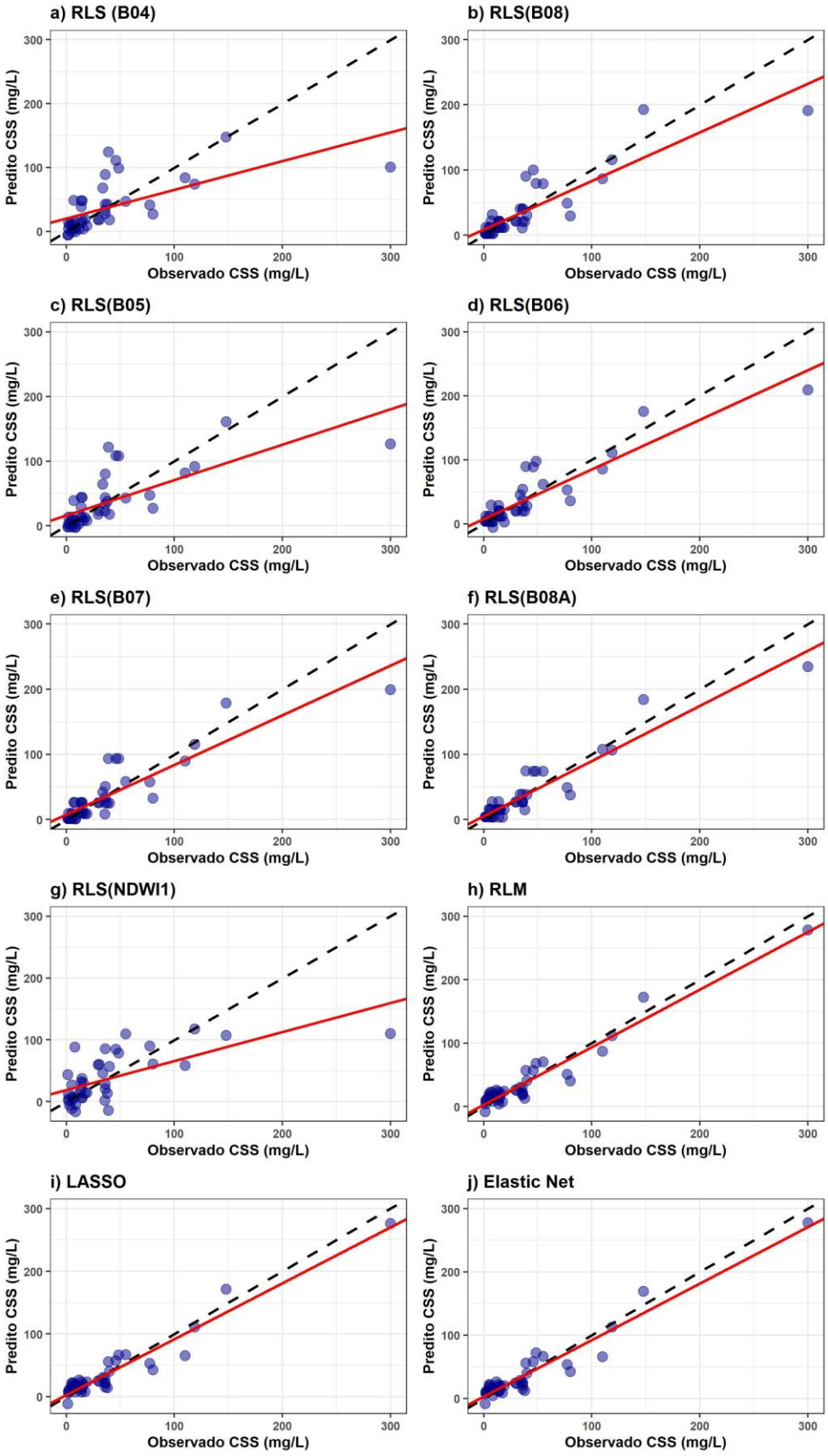


Figura 1.3. Valores observados e preditos por meio das variáveis preditoras utilizando o satélite MSI/Sentinel 2 para cada modelo utilizado.

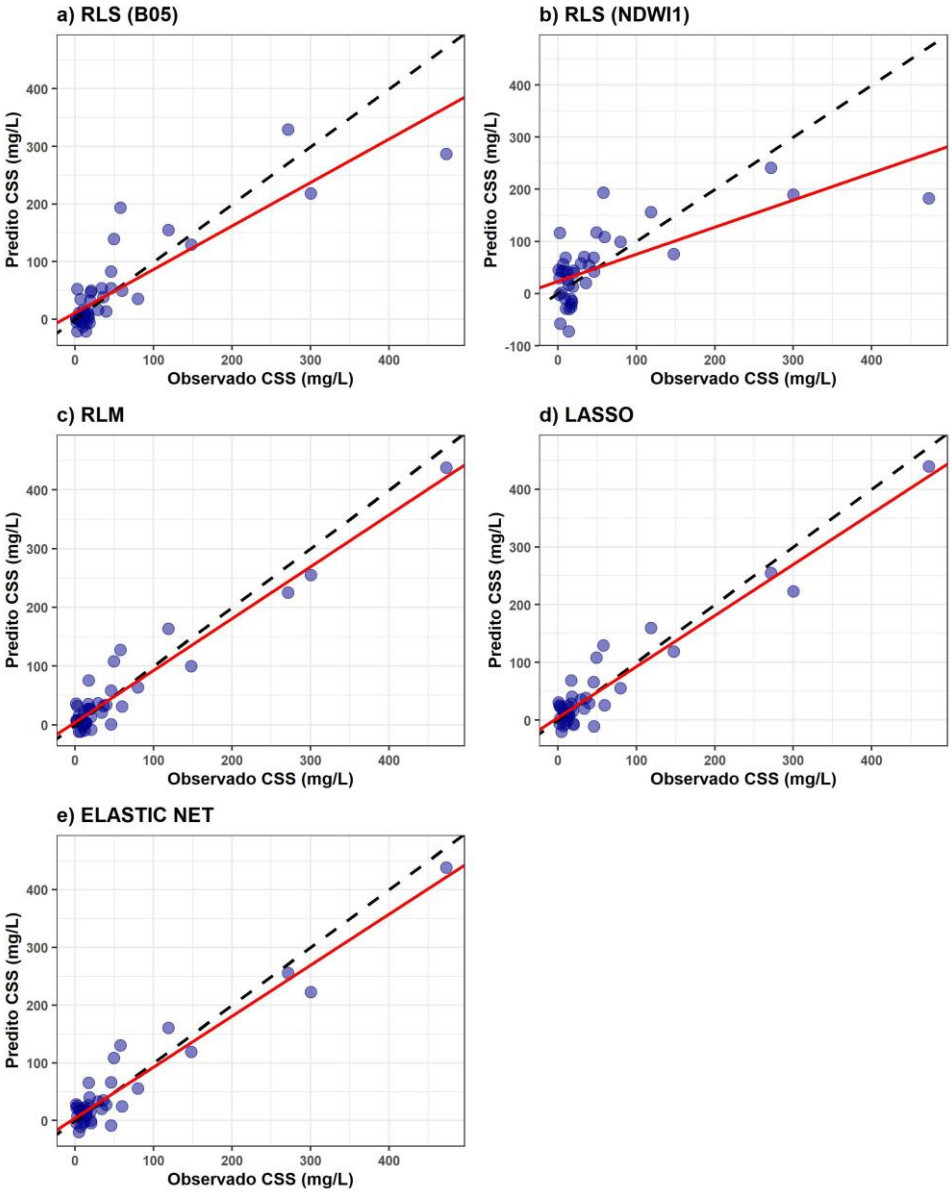


Figura 1.4. Valores observados e preditos obtidos por meio das variáveis preditoras utilizando o satélite OLI/Landsat 8 para cada modelo utilizado.

É possível notar nas Figuras 1.3 e 1.4 que os modelos que melhor se aproximaram da reta 1:1 também foram aqueles que apresentaram bom desempenho, como pode ser analisado na Tabela 1.6. Os modelos que utilizaram múltiplas variáveis estimaram valores negativos da CSS em alguns casos em que seu valor foi baixo (< 2 mg/L), o que não apresenta significado físico. Isso ocorreu principalmente por causa da banda do vermelho, que apresentou coeficiente de regressão (β_1) negativo, e em situação em que há pouco material em suspensão na água (períodos de estiagens), esta tende a apresentar maiores valores de refletância e maior absorção na região do infravermelho próximo (BARBOSA; NOVO; MARTINS, 2019).

Apenas o modelo de RLS utilizando a banda B08A do satélite MSI/Sentinel 2 apresentou todos os valores preditos positivos da CSS. Entretanto, como o período mais crítico para o monitoramento dos sedimentos se concentra na estação chuvosa (FROMANT et al., 2021), optou-se por gerar os mapas de fluxos de sedimentos empregando o modelo RLM a partir das variáveis preditoras derivadas do satélite MSI/Sentinel 2, o qual apresentou o melhor desempenho.

1.3.3. Mapas de fluxos de sedimentos

Foram gerados mapas de fluxos de sedimentos em suspensão ao longo da calha do rio principal da bacia hidrográfica do rio Doce. O período de análise baseou-se em duas etapas, sendo a primeira concentrada ao final 2015 e durante 2016, com intuito de avaliar a dinâmica dos sedimentos para o período posterior ao rompimento da barragem de Fundão, ocorrido em 05 de novembro de 2015. Na segunda etapa verificou-se a dinâmica da CSS ao longo dos anos de 2019 e 2020, para analisar se ocorreram modificações nas concentrações de sedimentos anos após o rompimento da barragem.

As Figuras 1.5 e 1.6 representam os fluxos da CSS em sete trechos mais próximos ao local de desague dos rejeitos da barragem de Fundão na calha do rio Doce, e em sete trechos mais distantes, respectivamente, incluindo a foz da bacia, entre 2015 e 2016. Os mapas de fluxo de sedimentos são apresentados em ordem cronológica, e a definição de cada trecho baseou-se nos locais em que foi possível a obtenção de imagens orbitais sem a presença de nuvens, desta forma, os trechos não estão em sequência.

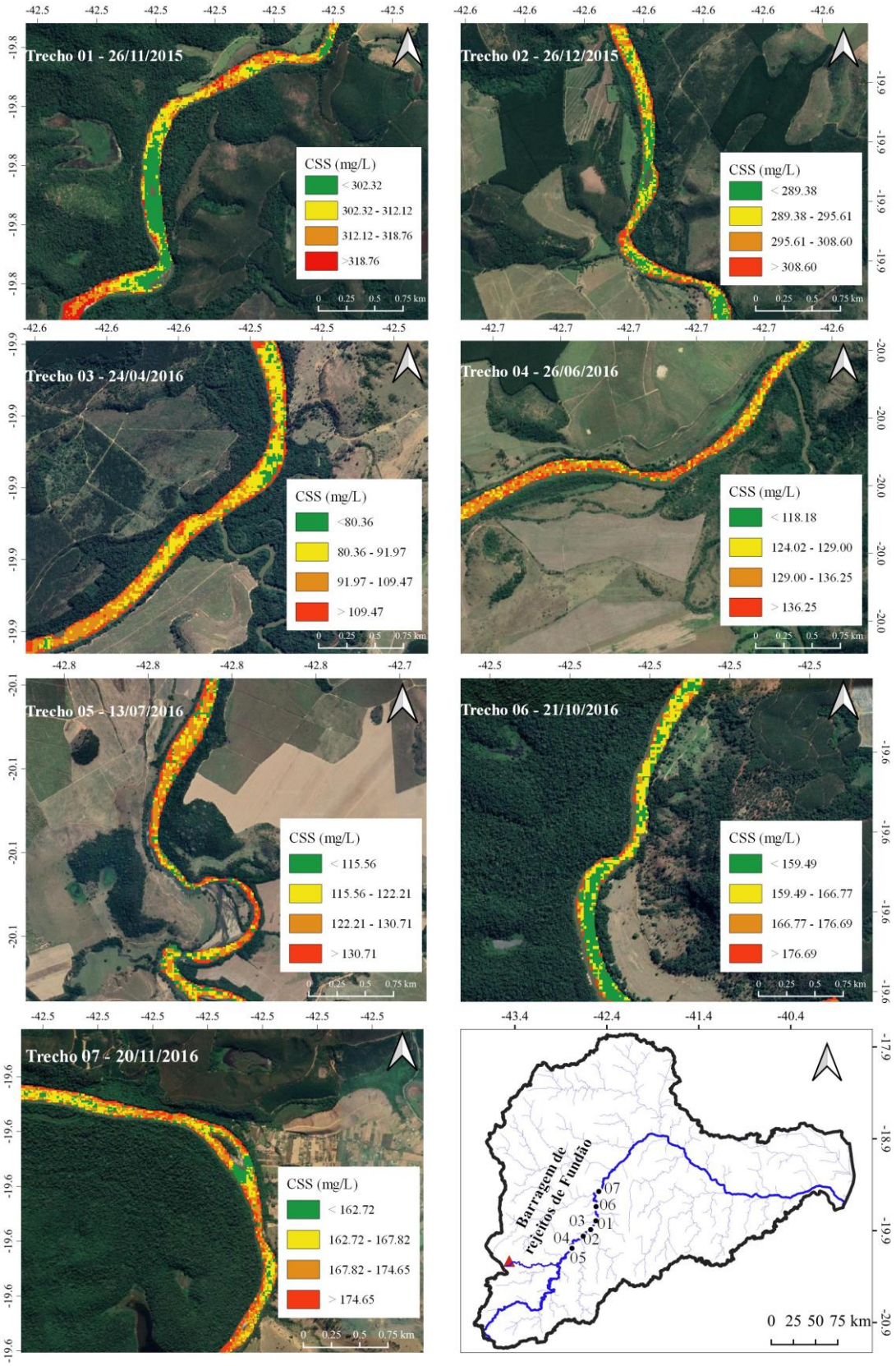


Figura 1.5. Dinâmica da concentração superficial de sedimentos (CSS) entre 2015 e 2016 estimada a partir de imagens do satélite MSI/Sentinel 2 para trechos da calha do rio Doce próximos ao local de desague dos rejeitos da barragem de Fundão.

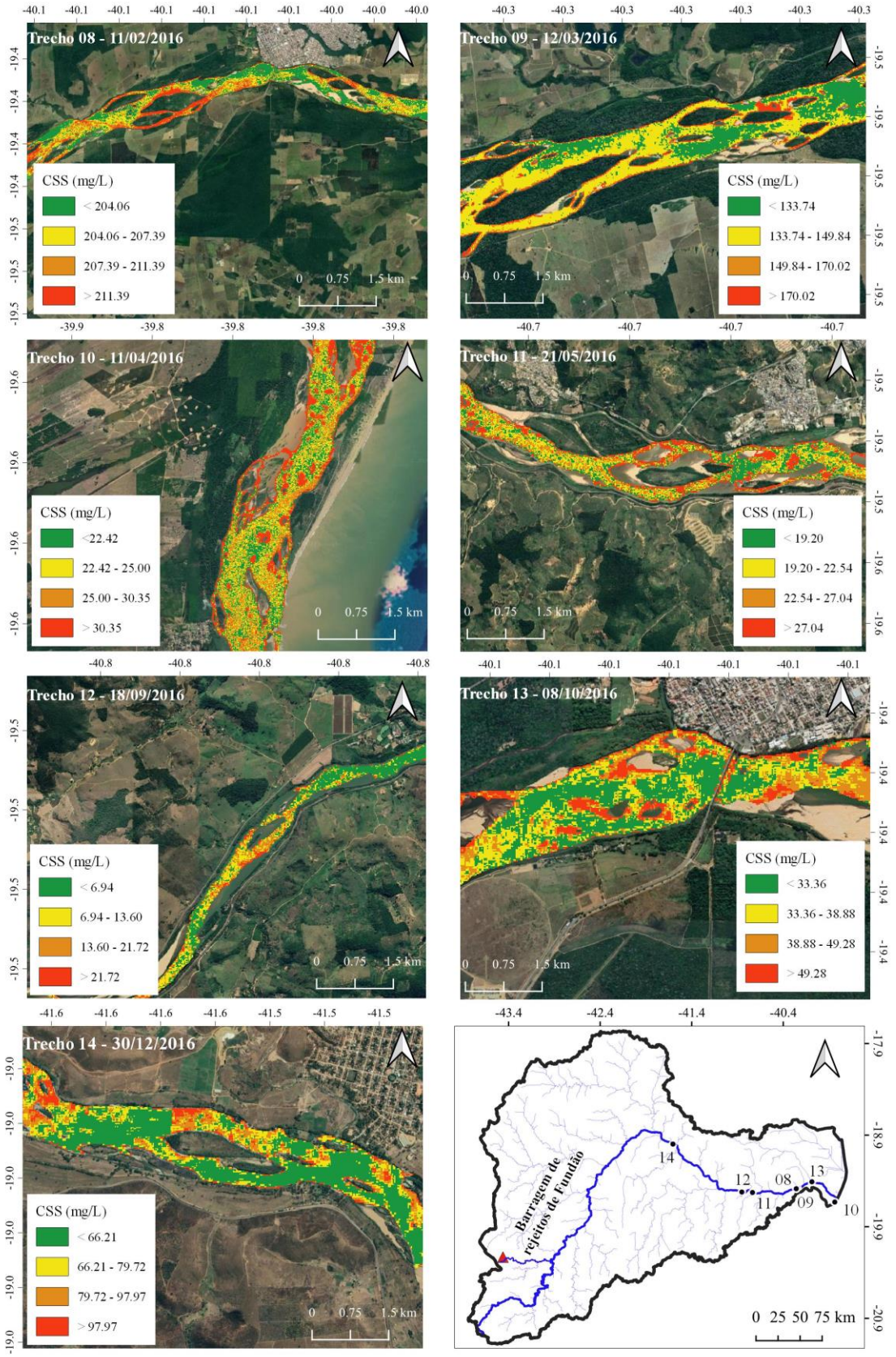


Figura 1.6. Dinâmica da concentração superficial de sedimentos (CSS) entre 2015 e 2016 estimada a partir de imagens do satélite MSI/Sentinel 2 para trechos da calha do rio Doce mais afastado do local de desague dos rejeitos da barragem de Fundão.

Pode-se verificar nas Figura 1.5 e 1.6 que os maiores valores da CSS estimados ocorreram nos meses subsequentes ao rompimento da barragem de Fundão. Nos trechos 1 e 2, com as imagens do satélite MSI/Sentinel 2 nas datas 26 de novembro de 2015 e 26 de dezembro de 2015, foram observados valores da CSS superiores a 318 mg/L e 308 mg/L, respectivamente.

Devido a presença de nuvens e da indisponibilidade de imagens, na região mais distante do local de deságue dos rejeitos da barragem de Fundão, foi possível estimar o fluxo de CSS apenas a partir de 11 de fevereiro de 2016, no trecho 8. Nesse local, o valor estimado da CSS foi superior a 211 mg/L.

Observa-se que os valores de CSS na região mais próxima a barragem de rejeitos de Fundão permaneceram elevados mesmo em períodos em que o volume pluviométrico, em geral, é mais baixo. Como pode ser observado nos trechos 4 e 5 em que foram estimados valores de superiores a 136 mg/L e 130 mg/L, em 26 de junho de 2016 e 13 de julho de 2016, respectivamente.

Nos trechos mais afastados do local de deságue dos rejeitos da barragem de Fundão os valores estimados foram relativamente menores, pois os rios mais largos favorecem a deposição do material em suspensão, já que as condições hidráulicas da coluna de água influenciam diretamente na deposição ou ressuspensão dos sedimentos (WILKES et al., 2019). Nos trechos 11 e 12 foram estimados valores superiores a 27 mg/L e 21 mg/L, em 21 de maio de 2016 e 18 de setembro de 2016, respectivamente.

Destaca-se, também, que os valores de CSS aumentam conforme a proximidade do período chuvoso na bacia hidrográfica do rio Doce, que se concentra entre outubro e março (ECOPLAN-LUME, 2010), sendo verificados nos trechos 07 e 14, em 20 de novembro de 2016 e 30 de dezembro de 2016, valores estimados superiores a 174 mg/L e 97 mg/L, respectivamente.

Na Figura 1.7 é possível analisar os valores estimados da CSS ao longo dos anos de 2019 e 2020. Os trechos selecionados estão localizados em locais próximos as estações sedimentométricas da ANA e Fundação Renova.

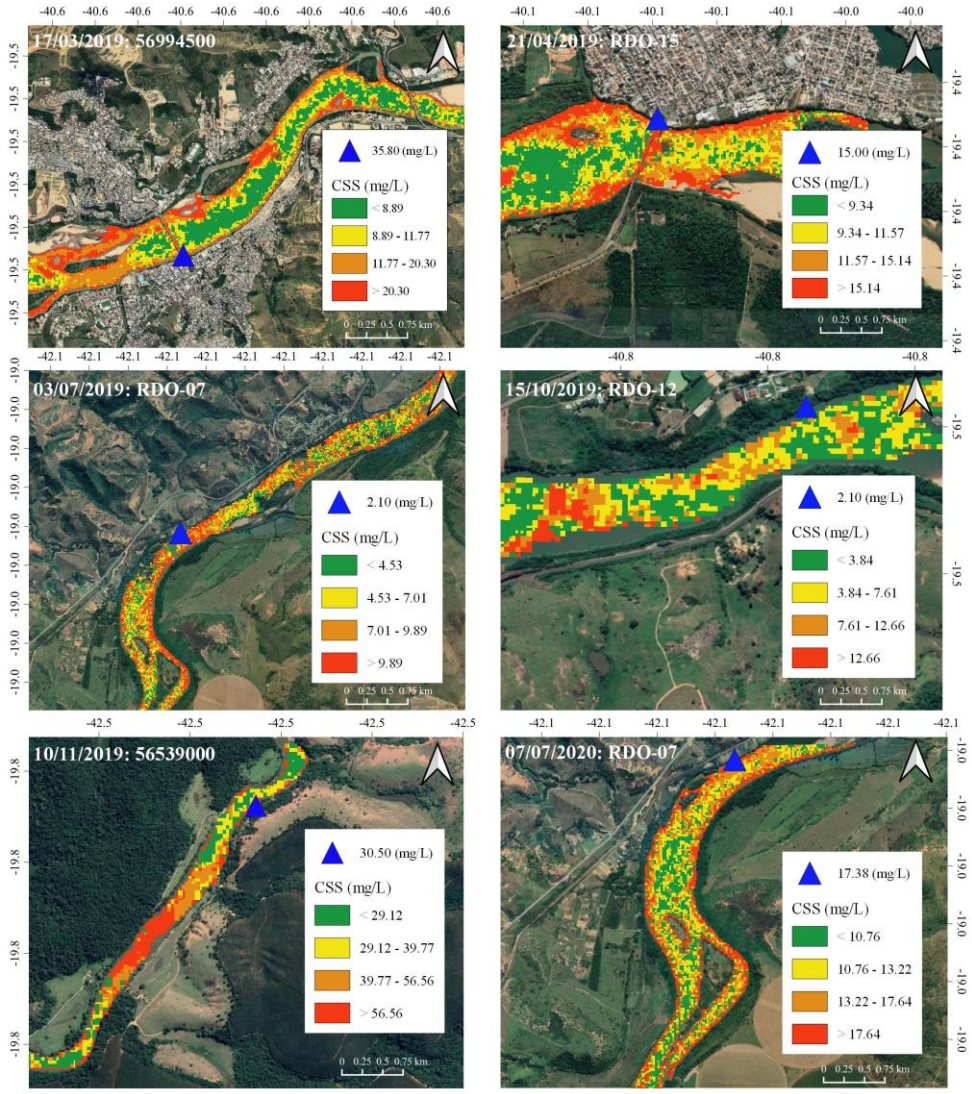


Figura 1.7. Valores estimados da CSS por meio de imagens do satélite MSI/Sentinel 2 e valor observado na respectiva estação sedimentométrica ao longo dos anos de 2019 e 2020, utilizando o satélite MSI/Sentinel 2.

É importante destacar na Figura 1.7 que na maioria das estações os valores estimados da CSS por meio das imagens orbitais entre 2019 e 2020 foram próximos aos dados observados nas estações, principalmente nos locais próximos de medição. Há grande variabilidade da CSS estimada ao longo da calha do rio, que apenas à algumas distâncias do ponto de coleta, apresentaram valores bastantes diferentes, como pode ser analisado na estação 56539000 em 10 de novembro de 2019, em que há trechos do rio com CSS superiores a 56 mg/L, enquanto o valor observado foi de 30,5 mg/L. Isso configura-se uma desvantagem ao se trabalhar com dados pontuais, pois representam apenas os trechos da calha do rio em que os sedimentos foram amostrados (HAMAAMIN et al., 2019).

Na Figura 1.7 é possível observar que os valores da CSS estimados são menores em comparação com aqueles estimados nos anos subsequentes ao rompimento da barragem de Fundão (Figuras 1.5 e 1.6). Isso demonstra que grande parte dos rejeitos que ficaram depositados na calha do rio após o acidente podem ter sido transportados até o oceano, dada a dinâmica da ressuspensão, transporte e deposição que envolve a sedimentologia fluvial (CARVALHO et al., 2000; POLETO, 2018). O que também pode ser analisado na Figura 1.8, que apresenta a CSS e os valores de vazão fluvial medidos nas estações da ANA e Fundação Renova entre 2015 e 2020, em sequência de instalação próxima da região de desague dos rejeitos da barragem de fundão até a foz da bacia hidrográfica do rio Doce.

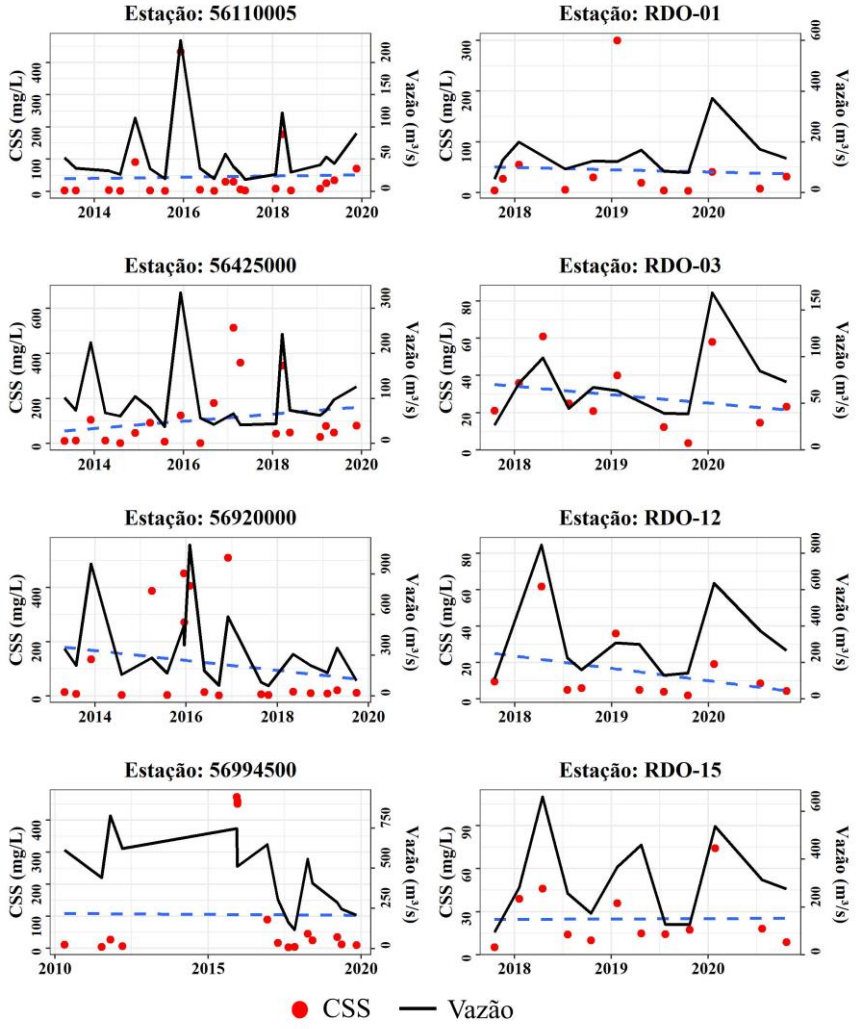


Figura 1.8. Valores observados da CSS e os valores de vazão fluvial medidos nas estações sedimentométricas da ANA e Fundação Renova da região mais próxima para a mais afastada do local de rompimento da barragem de Fundão.

As estações com códigos 55110005, 56425000, RDO-01 e RDO-03 apresentadas na Figura 1.8 estão localizadas próximas ao local de deságue dos rejeitos da barragem de Fundão, enquanto que as estações com códigos 56920000, 56994500, RDO-12 e RDO-15, mais distantes. As estações da ANA apresentam dados históricos mais longos, onde é possível notar que os maiores valores de sedimentos em suspensão, em geral, estão concentrados entre 2015 e 2016, com valores superiores a 400 mg/L. Em anos mais recentes, percebe-se certa estabilidade da CSS, com valores de próximos a 100 mg/L, conforme os valores de vazão fluvial mais elevados.

Como pode ser observado na Figura 1.8, em alguns casos a CSS foi proporcional ao aumento da vazão, entretanto, em grande parte não se observa relação linear, o que dificulta obter a CSS a partir dos dados de vazão. Isso faz com que a utilização de sensoriamento remoto orbital seja de grande auxílio para a gestão dos recursos hídricos na bacia hidrográfica do rio Doce.

Os dados observados da CSS demonstram que o monitoramento do fluxo de sedimentos por meio de sensoriamento remoto orbital teve boa representação da dinâmica dos sedimentos na calha do rio Doce. Os resultados obtidos no presente trabalho são corroborados por diversos outros estudos utilizando sensoriamento remoto orbital, nos quais foram observadas forte relação entre os dados observados da CSS com os dados estimados por satélites (MARTINEZ et al., 2009; PARK; LATRUBESSE, 2014, 2015; ESPINOZA-VILLAR et al., 2018; MARINHO et al., 2018; GALLAY et al., 2019; SABERIOON et al., 2020; ZAHIRI; MOLLAEE; ANSARI, 2020).

A grande vantagem da utilização de sensoriamento remoto orbital consiste na obtenção da variação da CSS ao longo da calha do rio e não somente em um único ponto, como é o caso das estações sedimentométricas. Além disso, nota-se que nem sempre é possível observar a relação entre a CSS e vazão fluvial, a qual é utilizada para gerar a curva chave de sedimentos (POLETO, 2018; HAMAAMIN et al., 2019).

Os mapas de fluxo de sedimentos permitem identificar os locais com maior concentração de sedimentos ao longo do rio. Satélites como o MSI/Sentinel 2 e o OLI/Landsat 8 trouxeram grandes avanços para o monitoramento da CSS (PETERSON et al., 2018; SABERIOON et al., 2020). No entanto, o monitoramento ainda é restrito a rios mais largos, como foi o caso da bacia hidrográfica do rio Doce, em que as medições ficaram concentradas ao longo da calha do rio principal.

A presença de nuvens também dificulta o monitoramento contínuo, principalmente no período chuvoso, o qual é o período mais crítico no monitoramento dos sedimentos

(FROMANT et al., 2021). Tanto que não foi possível a obtenção de imagens para janeiro de 2016 para incluir o mapa de fluxo de sedimentos (Figuras 1.5 e 1.6). Por isso, é imprescindível a utilização de diferentes satélites para o monitoramento dos recursos hídricos. Com o lançamento do Landsat 9, ocorrido em setembro de 2021, diminuiu-se o tempo de revisita da constelação desse satélite para oito dias (NASA, 2021) e, desta forma, a utilização integrada do MSI/Sentinel 2 com o OLI/Landsat 8 e OLI-2/Landsat 9 pode aumentar bastante a disponibilidade de imagens para o monitoramento.

1.4 CONCLUSÕES

Com base nos resultados obtidos no presente trabalho pode-se concluir que:

- É possível realizar o monitoramento da concentração superficial de sedimentos (CSS) utilizando sensoriamento remoto orbital na calha do rio principal da bacia hidrográfica do rio Doce, por meio da relação linear entre a refletância medida pelo sensor orbital e os dados observados da CSS;
- A banda do infravermelho próximo apresentou forte relação linear com a CSS, tanto utilizando o satélite MSI/Sentinel 2 (B08A) quanto o OLI/Landsat 8 (B05);
- As bandas do visível e infravermelho próximo (VNIRs) com 20 m de resolução espacial do satélite MSI/Sentinel 2 apresentaram boa relação linear no monitoramento da CSS, evidenciando o potencial desse satélite para o monitoramento dos recursos hídricos;
- Dentre os modelos de regressão linear que utilizam múltiplas variáveis, tanto a regressão linear múltipla quando a regressão LASSO e a regressão *Elastic Net* apresentaram bom desempenho para a predição dos sedimentos em suspensão, principalmente utilizando o Satélite MSI/Sentinel 2. Entretanto, estas últimas facilitam na definição do conjunto ótimo de variáveis;
- Os mapas de fluxos de sedimentos indicam redução da CSS na calha do rio Doce em anos mais recentes, o que pode ser indicativo de que parte do material oriundo do rompimento da barragem de rejeitos de Fundão pode ter sido carreado pelos processos de ressuspensão e transporte de sedimentos.

REFERÊNCIAS

AFAN, H. A. et al. Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction. **Journal of Hydrology**, v. 541, p. 902–913, 1 out. 2016.

AICH, V.; ZIMMERMANN, A.; ELSENBEER, H. Quantification and interpretation of suspended-sediment discharge hysteresis patterns: How much data do we need? **CATENA**, v. 122, p. 120–129, 1 nov. 2014.

AL-MUKHTAR, M. Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. **Environmental Monitoring and Assessment**, v. 191, n. 11, p. 673, 25 nov. 2019.

AL-MUKHTAR, M.; AL-YASEEN, F. Modeling Water Quality Parameters Using Data-Driven Models, a Case Study Abu-Ziriq Marsh in South of Iraq. **Hydrology**, v. 6, n. 1, p. 24, 17 mar. 2019.

ALTHOFF, D.; RODRIGUES, L. N. Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. **Journal of Hydrology**, v. 600, p. 126674, 1 set. 2021.

ALIN, A. Multicollinearity. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 2, n. 3, p. 370–374, 2010.

ANDRZEJ URBANSKI, J. et al. Application of Landsat 8 imagery to regional-scale assessment of lake water quality**. International Journal of Applied Earth Observation and Geoinformation**, v. 51, p. 28–36, 1 set. 2016

BARBOSA, C.; NOVO, E.; MARTINS, V. **Introdução ao Sensoriamento Remoto de sistemas aquáticos**. 1. ed. São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 161p. 2019., 2019.

BHANGALE, U. et al. Analysis of Surface Water Resources Using Sentinel-2 Imagery. **Procedia Computer Science**, v. 171, p. 2645–2654, 1 jan. 2020.

BIRTH, G. S.; MCVEY, G. R. Measuring the Color of Growing Turf with a Reflectance Spectrophotometer1. **Agronomy Journal**, v. 60, n. 6, p. 640–643, 1 nov. 1968.

BOGGS, S. **Principles of Sedimentology and Stratigraphy**. 4. ed. Upper Saddle River: Pearson Prentice Hall, 622 p., 2006.

CAMARGO, ângelo P. de; SENTELHAS, P. C. Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no Estado de São Paulo. **Revista Brasileira de Agrometeorologia**, v. 5, n. 1, p. 89–87, 1997.

CAO, L. et al. Factors controlling discharge-suspended sediment hysteresis in karst basins, southwest China: Implications for sediment management. **Journal of Hydrology**, v. 594, p. 125792, 1 mar. 2021.

CARVALHO, N. de O. et al. **Guia de práticas sedimentométricas**. Brasília: ANEEL, 2000.

CBH-DOCE. **A bacia hidrográfica do Rio Doce**. Disponível em: <http://www.cbhdoce.org.br/institucional/a-bacia>. Acesso em: 5 out. 2018.

DE MOL, C.; DE VITO, E.; ROSASCO, L. Elastic-net regularization in learning theory. **Journal of Complexity**, v. 25, n. 2, p. 201–230, 1 abr. 2009.

ECOPLAN-LUME. **Plano Integrado de Recursos Hídricos da Bacia Hidrográfica do Rio Doce - PIRH Bacia do Rio Doce**. 1. ed. Belo Horizonte: CONSÓRCIO ECOPLAN-LUME, 2010.

ESA. **Spatial Resolutions Sentinel-2 MSI**. Disponível em: <https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/spatial>. Acesso em: 7 out. 2018.

ESA. **Copernicus Open Access Hub**. Disponível em: <https://scihub.copernicus.eu/twiki/do/view/SciHubWebPortal/WebHome#dias-box>. Acesso em: 11 jan. 2022.

ESPINOZA-VILLAR, R. et al. Spatio-temporal monitoring of suspended sediments in the Solimões River (2000–2014). **Comptes Rendus Geoscience**, v. 350, n. 1–2, p. 4–12, 1 jan. 2018.

FEYISA, G. L. et al. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. **Remote Sensing of Environment**, v. 140, p. 23–35, 1 jan. 2014.

FOX, J., WEISBERG, S. An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage, 2019. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

FROMANT, G. et al. Suspended sediment concentration field quantified from a calibrated MultiBeam EchoSounder. **Applied Acoustics**, v. 180, p. 108107, 1 set. 2021.

GALLAY, M. et al. Assessing Orinoco river sediment discharge trend using MODIS satellite images. **Journal of South American Earth Sciences**, v. 91, p. 320–331, 1 abr. 2019.

GAO, B. C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. **Remote Sensing of Environment**, v. 58, n. 3, p. 257–266, 1 dez. 1996.

GEE. **Google Earth Engine**. Disponível em: <https://earthengine.google.com/>. Acesso em: 11 jan. 2022.

GERACE, A. D.; SCHOTT, J. R.; NEVINS, R. Increased potential to monitor water quality in the near-shore environment with Landsat's next-generation satellite. **Journal of Applied Remote Sensing**, v. 7, n. 1, p. 073558, 22 maio 2013.

GUPTA, H. V. et al. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. **Journal of Hydrology**, v. 377, n. 1–2, p. 80–91, 20 out. 2009.

HADDAD, K. et al. Applicability of Monte Carlo cross validation technique for model development and validation using generalised least squares regression. **Journal of Hydrology**, v. 482, p. 119–128, 4 mar. 2013.

HADDADCHI, A.; HICKS, M. Interpreting event-based suspended sediment concentration and flow hysteresis patterns. **Journal of Soils and Sediments**, v. 21, p. 592–612, 2021.

HAIR, J. F. et al. Análise de Regressão Múltipla. In: HAIR, J. F. et al. (Ed.). **Análise Multivariada de Dados**. 6. ed. Porto alegre: Bookman: Bookman, 2009. p. 149–220.

HALLAK, R.; PEREIRA FILHO, A. J. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS. **Revista Brasileira de Meteorologia**, v. 26, n. 4, p. 591–608, 2011.

HAMAAMIN, Y. A. et al. Evaluation of neuro-fuzzy and Bayesian techniques in estimating suspended sediment loads. **Sustainable Water Resources Management**, v. 5, n. 2, p. 639–654, 2019.

HARMEL, T. et al. Sunglint correction of the Multi-Spectral Instrument (MSI)-SENTINEL-2 imagery over inland and sea waters from SWIR bands. **Remote Sensing of Environment**, v. 204, n. January, p. 308–321, 2018.

HIDROWEB. **Séries Históricas de Estações**. Disponível em: <https://www.snirh.gov.br/hidroweb/serieshistoricas>. Acesso em: 11 jan. 2022.

IBGE. **Downloads: Geociências**. Disponível em: <https://www.ibge.gov.br/geociencias/downloads-geociencias.html>. Acesso em: 7 jan. 2022.

JALLY, S. K.; MISHRA, A. K.; BALABANTARAY, S. Retrieval of suspended sediment concentration of the Chilika Lake, India using Landsat-8 OLI satellite data. **Environmental Earth Sciences**, v. 80, n. 8, p. 1–18, 1 abr. 2021.

JAELANI, L. M. et al. Estimation of Total Suspended Sediment and Chlorophyll-A Concentration from Landsat 8-Oli: The Effect of Atmospher and Retrieval Algorithm. **IPTEK The Journal for Technology and Science**, v. 27, n. 1, 22 abr. 2016.

JAMES, G. et al. **An Introduction to Statistical Learning**. 2. ed. New York: Springer, 2021.

KNAEPS, E. et al. A SWIR based algorithm to retrieve total suspended matter in extremely turbid waters. **Remote Sensing of Environment**, v. 168, p. 66–79, 2015.

KUMAR, D. et al. Daily suspended sediment simulation using machine learning approach. **Catena**, v. 138, p. 77–90, 2016.

LACAUX, J. P. et al. Classification of ponds from high-spatial resolution remote sensing: Application to Rift Valley Fever epidemics in Senegal. **Remote Sensing of Environment**, v. 106, n. 1, p. 66–74, 15 jan. 2007.

LEGATES, D. R.; MCCABE, G. J. Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. **Water Resources Research**, v. 35, n. 1, p. 233–241, 1 jan. 1999.

LI, P. et al. Soil erosion rates assessed by RUSLE and PESERA for a Chinese Loess Plateau catchment under land-cover changes. **Earth Surface Processes and Landforms**, v. 45, n. 3, p. 707–722, 15 mar. 2020.

LI, T. et al. Driving forces and their contribution to the recent decrease in sediment flux to ocean of major rivers in China. **Science of The Total Environment**, v. 634, p. 534–541, 1 set. 2018.

LI, Z. et al. Impacts of land use change and climate variability on hydrology in an agricultural catchment on the Loess Plateau of China. **Journal of Hydrology**, v. 377, n. 1–2, p. 35–42, 20 out. 2009.

LIMA, J. E. F. W. et al. Suspended sediment fluxes in the large river basins of Brazil. In: Sediment Budgets I, Seventh IAHS Scientific Assembly, Foz do Iguaçu. **Anais**... Foz do Iguaçu: IAHS, 2005.

LYRA, B. U.; RIGO, D. Deforestation impact on discharge regime in the Doce River Basin. **Revista Ambiente & Água**, v. 14, n. 4, 15 jul. 2019.

KUHN, M. **caret: Classification and Regression Training**. R package version 6.0-89, 2021.

https://CRAN.R-project.org/package=caret.

MALIK, A.; KUMAR, A.; PIRI, J. Daily suspended sediment concentration simulation using hydrological data of Pranhita River Basin, India. **Computers and Electronics in Agriculture**, v. 138, p. 20–28, 2017.

MAPBIOMAS. **Mapas e dados**. Disponível em: <https://mapbiomas.org/colecoes-mapbiomas-1?cama_set_language=pt-BR>. Acesso em: 10 jan. 2022.

MARINHO, T. et al. Suspended Sediment Variability at the Solimões and Negro Confluence between May 2013 and February 2014. **Geosciences**, v. 8, n. 7, p. 265, 19 jul. 2018.

MARINHO, R. R. et al. Spatiotemporal Dynamics of Suspended Sediments in the Negro River, Amazon Basin, from In Situ and Sentinel-2 Remote Sensing Data. **International Journal of Geo-Information** 2021, Vol. 10, Page 86, v. 10, n. 2, p. 86, 19 fev. 2021.

MARTINEZ, J. M. et al. Increase in suspended sediment discharge of the Amazon River assessed by monitoring network and satellite data. **Catena**, v. 79, n. 3, p. 257–264, 2009.

MUKHERJEE, N. R.; SAMUEL, C. Assessment of the Temporal Variations of Surface Water Bodies in and around Chennai using Landsat Imagery. **Indian Journal of Science and Technology**, v. 9, n. 18, p. 1–7, 14 maio 2016.

NAGUETTINI, M.; PINTO, E. J. A. **Hidrologia Estatística**. 1. ed. Belo Horizonte: CPRM, 2007.

NASA. **MODIS Land Science Team**. Disponível em: <https://modis-land.gsfc.nasa.gov/>. Acesso em: 7 out. 2018a.

NASA. **History Landsat Science**. Disponível em: <https://landsat.gsfc.nasa.gov/about/history/>. Acesso em: 29 jan. 2018b.

NASA. **Landsat 9 Landsat Science**. Disponível em: <https://landsat.gsfc.nasa.gov/satellites/landsat-9/>. Acesso em: 11 jan. 2022.

NAVRATIL, O. et al. Global uncertainty analysis of suspended sediment monitoring using turbidimeter in a small mountainous river catchment. **Journal of Hydrology**, v. 398, n. 3–4, p. 246–259, 24 fev. 2011.

PARK, E.; LATRUBESSE, E. M. Modeling suspended sediment distribution patterns of the Amazon River using MODIS data. **Remote Sensing of Environment**, v. 147, p. 232–242, 2014.

PARK, E.; LATRUBESSE, E. M. Surface water types and sediment distribution patterns at the confluence ofmega rivers: The Solim~ oes-Amazon and Negro Rivers junction. **Water Resources Research**, v. 51, p. 6197–6213, 2015.

PEREIRA, H. R. et al. On the performance of three indices of agreement: an easy-to-use r-code for calculating the Willmott indices. **Bragantia**, v. 77, n. 2, p. 394–403, 22 mar. 2018.

PETERSON, K. T. et al. Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine. **Remote Sensing**, v. 10, n. 10, p. 1–17, 2018.

POLETO, C. **Sedimentologia Fluvial: Estudos e Tecnicas**. 2. ed. Porto Alegre: ABRH, 2018.

QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System**. Open Source Geospatial Foundation Project. http://qgis.osgeo.org, 2021.

R Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

RENOVA. **Portal de Monitoramento rio Doce**. Disponível em: <https://portal-de-monitoramento-rio-doce-fundacaorenova.hub.arcgis.com/pages/pa-download>. Acesso em: 11 jan. 2022.

RESTREPO, J. D.; ESCOBAR, H. A. Sediment load trends in the Magdalena River basin (1980–2010): Anthropogenic and climate-induced causes. **Geomorphology**, v. 302, p. 76–91, 2018.

RITTER, A.; MUÑOZ-CARPENA, R. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. **Journal of Hydrology**, v. 480, p. 33–45, 14 fev. 2013.

SABERIOON, M. et al. Chlorophyll-a and total suspended solids retrieval and mapping using Sentinel-2A and machine learning for inland waters. **Ecological Indicators**, v. 113, p. 106236, 1 jun. 2020.

SANTOS, W. D. O. et al. Revista Brasileira de Geografia Física. **Revista Brasileira de Geografia Física**, v. 7, n. 3, p. 434–441, 2014.

VILLAR, R. E. et al. A study of sediment transport in the Madeira River, Brazil, using MODIS remote-sensing images. **Journal of South American Earth Sciences**, v. 44, p. 45–54, 2013.

WARREN, M. A. et al. Assessment of atmospheric correction algorithms for the Sentinel-2A MultiSpectral Imager over coastal and inland waters. **Remote Sensing of Environment**, v. 225, p. 267–289, 1 maio 2019.

WEI, T., SIMKO, V. **R package 'corrplot'**: Visualization of a Correlation Matrix (Version 0.92), 2021. Available from https://github.com/taiyun/corrplot

WILKES, M. A. et al. Physical and biological controls on fine sediment transport and storage in rivers. **Wiley Interdisciplinary Reviews: Water**, v. 6, n. 2, p. e1331, 1 mar. 2019.

XU, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. **https://doi.org/10.1080/01431160600589179**, v. 27, n. 14, p. 3025–3033, 20 jul. 2007.

YEPEZ, S. et al. Retrieval of suspended sediment concentrations using Landsat-8 OLI satellite images in the Orinoco River (Venezuela). **Comptes Rendus Geoscience**, v. 350, n. 1–2, p. 20–30, 1 jan. 2018.

ZAHIRI, J.; MOLLAEE, Z.; ANSARI, M. R. Estimation of Suspended Sediment Concentration by M5 Model Tree Based on Hydrological and Moderate Resolution Imaging Spectroradiometer (MODIS) Data. **Water Resources Management**, v. 34, n. 12, p. 3725–3737, 1 set. 2020.

ZAMBRANO-BIGIARINI, M. **hydroGOF:** Goodness-of-fit functions for comparison of simulated and observed hydrological time seriesR package version 0.4-0. 2020.

ZHAO, J. et al. Remote Sensing Evaluation of Total Suspended Solids Dynamic with Markov Model: A Case Study of Inland Reservoir across Administrative Boundary in South China. **Sensors** 2020, Vol. 20, Page 6911, v. 20, n. 23, p. 6911, 3 dez. 2020.

ZHAO, Y. et al. Analysis of changes in characteristics of flood and sediment yield in typical basins of the Yellow River under extreme rainfall events. **CATENA**, v. 177, p. 31–40, 1 jun. 2019.

CAPÍTULO 2:

Modelagem da concentração superficial de sedimentos na bacia hidrográfica do rio Doce com base em aprendizado de máquina

RESUMO: As medições dos sedimentos são bastante trabalhosas e onerosas, sendo fundamental a utilização de técnicas que possam obter essa informação a partir de variáveis mais simples de serem medidas. Desta forma, o objetivo deste trabalho foi utilizar modelos baseados em aprendizado de máquina para a predição da concentração superficial de sedimentos (CSS) na bacia hidrográfica do rio Doce. Os dados de CSS observados que foram utilizados na modelagem são valores médios da seção transversal medidos em sete estações sedimentométricas da Agência Nacional de Águas e Saneamento Básico (ANA) localizadas na calha do rio Doce. Foram utilizadas 62 variáveis preditoras derivadas das informações de declividade, pedologia, uso e cobertura da terra, precipitação, vazão fluvial, velocidade fluvial, evapotranspiração real, escoamento superficial, umidade do solo, temperatura e *normalized difference vegetation index* (NDVI). Com o intuito de reduzir o número de variáveis preditivas foram empregados os métodos de seleção de variáveis Boruta e *Recursive Feature Elimination* (RFE). Para a predição dos dados de CSS foram aplicados os algoritmos *Random Forest* (RF), *Cubist, Support Vector Machines* (SVMs), *Extreme Gradient Boosting Machine* (XGBoost) e regressão *Least Absolute Shrinkage and Selection Operator* (LASSO). Bons resultados foram obtidos com a utilização de algoritmos de aprendizado de máquina para a predição da CSS na bacia hidrográfica do rio Doce, com destaque para os modelos Cubist e XGBoost, que apresentaram o menor erro de predição e métricas de eficiência mais elevadas. As variáveis preditivas mais importantes na modelagem da CSS foram as vazões fluviais diárias da data da coleta dos sedimentos e as vazões defasadas no tempo. A precipitação média diária acumulada também foi importante na modelagem dos sedimentos. Com a realização do trabalho comprovou-se que os modelos de aprendizado de máquina podem ser de grande auxílio para o monitoramento dos sedimentos e servir como ferramenta para entender a dinâmica da produção de sedimentos na bacia hidrográfica do rio do Doce ao longo do tempo.

Palavras-chaves: modelagem hidrossedimentológica, monitoramento dos sedimentos, aprendizado supervisionado.

2.1. INTRODUÇÃO

O monitoramento da concentração de sedimentos em rios e reservatórios é fundamental para planejamento e gestão dos recursos hídricos, pois os sedimentos estão relacionados a problemas com a qualidade da água, assoreamento de rios e reservatórios, navegabilidade, degradação no ambiente aquático e mal funcionamento de usinas hidroelétricas (SHIAU; CHEN, 2015; KAVEH; DUC BUI; RUTSCHMANN, 2017; TAVAKOLI TARGHI; ABBASZADEH; ARABASADI, 2017; AL-MUKHTAR, 2019; FROMANT et al., 2021). A medição em campo da concentração de sedimentos é bastante trabalhosa, demandando tempo e recursos financeiros. Desta forma, é difícil obter uma base de dados detalhada das medições da CSS, sendo que na maioria dos casos as medições não são contínuas e contemplam apenas um curto período de tempo (MALIK; KUMAR; PIRI, 2017; AL-MUKHTAR; AL-YASEEN, 2019).

Em contraste, informações como a vazão fluvial estão disponíveis mais facilmente e em diferentes escalas de tempo, sendo possível a obtenção de dados horários ou, até mesmo, em intervalos mais curtos. Neste contexto, é de grande importância a obtenção de modelos que possam predizer a concentração de sedimentos a partir outras variáveis, mais simples de serem obtidas (AL-MUKHTAR, 2019). Diversos trabalhos recentes que estudaram o fenômeno de histerese da concentração superficial de sedimentos (CSS) apontam que há uma defasagem do aumento da CSS em comparação com o aumento da vazão fluvial, no entanto, é possível observar a relação entre as variáveis (HAMSHAW et al., 2018; KEESSTRA et al., 2019; MALUTTA et al., 2020; CAO et al., 2021; HADDADCHI; HICKS, 2021).

A inter-relação de fatores físicos e climáticos fazem da dinâmica dos sedimentos em rios não somente difícil de ser entendida, mas também de ser simulada (BHARTI et al., 2017). Isso ocorre, principalmente, por causa do seu comportamento não estacionário e alta variabilidade. A predição da concentração de sedimentos requer modelos que possam ter bom desempenho mesmo com dados faltantes e relação não linear com as variáveis explicativas (MALIK; KUMAR; PIRI, 2017). Devido a essa complexidade, as técnicas utilizadas para modelar esse fenômeno tem apresentado pouca capacidade preditiva (MUSTAFA, 2016) e não há uma aceitação universal dos modelos propostos (LAFDANI; NIA; AHMADI, 2013; MALIK; KUMAR; PIRI, 2017).

Diversas abordagens foram utilizadas para estudar a dinâmica dos sedimentos em bacias hidrográficas, como exemplos, os modelos USLE (*Universal Soil Loss Equation*) e RUSLE (*Revised Universal Soil Loss*) (MAGESH; CHANDRASEKAR, 2016; EFTHIMIOU;

LYKOUDI; KARAVITIS, 2017). Destaca-se, também, os modelos SWAT (*Soil and Water Assessment Tool*) (YESUF et al., 2015; RICCI; GIROLAMO, 2018) e WEPP (*Watershed Erosion Prediction Project*) (SRIVASTAVA et al., 2019). Dentre esses, o SWAT tem demostrado bons resultados na simulação da concentração superficial e transporte de sedimentos. No entanto, demanda grande quantidade de dados que, mesmo em bacias intensamente monitoradas, muitas vezes não há informações suficientes para sua calibração (BHARTI et al., 2017). Outro método utilizado para a estimativa da CSS é o uso da curva chave de sedimentos em suspensão, que consiste em um modelo de regressão na forma exponencial que relaciona à CSS com a vazão fluvial. Porém, também apresenta limitações em modelar o comportamento não estacionário desse processo (HAMAAMIN et al., 2019).

Modelos baseados em aprendizado de máquina e novos métodos computacionais trouxeram novas perspectivas para a modelagem da concentração de sedimentos (BHATTACHARYA; PRICE; SOLOMATINE, 2007; AFAN et al., 2016). Esses modelos têm demostrado boa capacidade para trabalhar a alta complexidade, dinamismo e não estacionariedade dos dados de sedimentos (NOURANI et al., 2014). Por serem modelos empíricos, a modelagem é feita com base na relação das informações de entrada e saída dos modelos e, por essa razão, tendem a serem mais acurados, pois dependem somente da base de dados (OLYAIE et al., 2015). No entanto, as equações matemáticas derivadas da análise dificilmente podem ser relacionadas com os processos físicos dos fenômenos estudados, sendo considerados, desta forma, como modelos caixa preta (CHEN; CHAU, 2016; ÖZGER; KABATAŞ, 2015).

Dentre os modelos baseados em aprendizado de máquina aplicados em hidrossedimentologia, destacam-se as *Artificial Neural Networks* (ANNs) (BUYUKYILDIZ; KUMCU, 2017; CHEN; CHAU, 2016; OLYAIE *et al.*, 2015; RAMEZANI; NIKOO, 2015; TAO; KESHTEGAR; YASEEN, 2019; YAWAR *et al.*, 2019), *Adaptive Neuro-Fuzzy Inference Systems* (ANFIS) (KISI; ZOUNEMAT-KERMANI, 2016; HAMAAMIN et al., 2019; KISI; MUNDHER, 2019; SAMET et al., 2019), Random Forest (RF) (CHEN et al., 2018; PETERSON et al., 2018; UMAR; RHOADS; GREENBERG, 2018; AL-MUKHTAR, 2019), *Support Vector Machines* (SVMs) (NOURANI; ALIZADEH; ROUSHANGAR, 2016; RASHIDI; VAFAKHAH, 2016; BUYUKYILDIZ; KUMCU, 2017; HIMANSHU; PANDEY; YADAV, 2017; RAHGOSHAY et al., 2019). O *Extreme Gradient Boosting* (XGBoost), o qual baseia-se em árvores de decisão, também tem apresentado excelentes resultados (NI et al., 2020), porém sua aplicação para a predição da concentração de sedimentos ainda é pouco explorada.

O monitoramento da CSS na calha do rio Doce é feito em poucas estações sedimentométricas. As campanhas de campo, em geral, são programadas para ocorrerem trimestralmente, no entanto, muitas estações apresentam dados faltantes, resultando em poucas informações para estudos mais detalhados sobre a dinâmica da produção de sedimentos na bacia hidrográfica do rio Doce. Portanto, a aplicação de modelos baseados em aprendizado de máquina pode resultar em melhorias no monitoramento da CSS, pois focam em variáveis preditoras mais fáceis de serem obtidas, como o regime de vazões e de precipitação. Neste contexto, o objetivo deste trabalho foi utilizar modelos baseados em aprendizado de máquina para a predição da concentração superficial de sedimentos e verificar as principais variáveis preditivas que influenciam na modelagem dos sedimentos na bacia hidrográfica do rio Doce.

2.2. MATERIAL E MÉTODOS

2.2.1 Região de estudo

A região de estudo é a bacia hidrográfica do rio Doce, localizada na região Sudeste do Brasil (Figura 2.1), com foco na calha principal do rio Doce, onde estão instaladas a maior parte das estações sedimentométricas da bacia. A bacia possui área de drenagem de, aproximadamente, 86,715 km², sendo 86% no estado de Minas Gerais e 14% no estado do Espirito Santo. O rio Doce tem extensão de 879 km e suas nascentes estão localizadas no estado de Minas Gerais, nas serras da Mantiqueira e do Espinhaço, e sua foz no oceano Atlântico na localidade de Vila Resende, no município de Linhares, Espirito Santo. O bioma predominante na região de estudo consiste na mata atlântica (98%) e cerrado (CBH-DOCE, 2018). A População da bacia é de cerca de 3,5 milhões de habitantes e as principais atividades econômicas são: mineração, siderurgia, silvicultura e agropecuária. (ELESBON et al., 2015).

A bacia hidrográfica do rio Doce é bastante antropizada com altas taxas de produção de sedimentos, em que grande parte dos usos da terra são destinados à agropecuária (63,5%) e floresta plantada e nativa (32,4%) (MAPBIOMAS, 2022). Os climas predominante na bacia de acordo com a classificação de Köppen são o clima temperado úmido com inverno seco e verão quente (Cwa) e clima tropical de savana com estação seca de inverno (Aw), com temperatura média anual de 18 ºC (ALVARES et al., 2013).

A precipitação média anual varia entre 836 mm a 1664 mm, com semestre chuvoso entre os meses de outubro a abril. As vazões máximas ocorrem nos meses de dezembro, janeiro e março, e as vazões mínimas, nos meses de agosto e setembro (ECOPLAN-LUME, 2010). Os

solos predominantes são o Latossolo Vermelho Amarelo (LVA), Argissolo Vermelho (PVe) e Cambissolos Háplicos (CXbe) (IBGE, 2019).

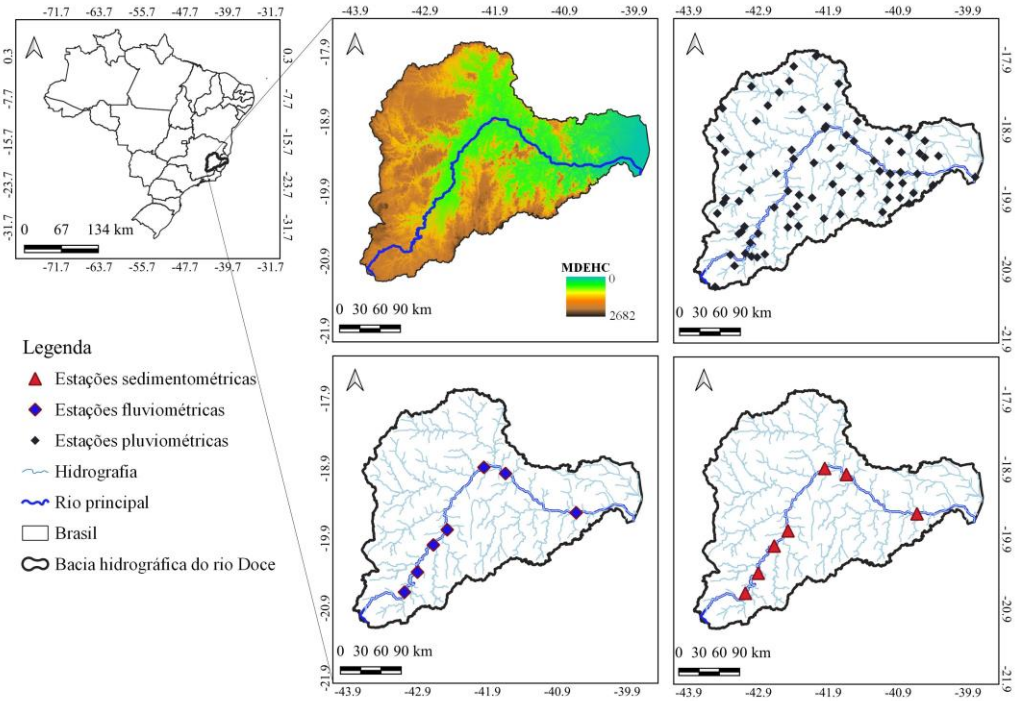


Figura 2.1 Localização da bacia hidrográfica do rio Doce, hidrografia, Modelo Digital de Elevação Hidrograficamente Condicionado (MDEHC), localização das estações sedimentométricas, pluviométricas e fluviométricas da Agência Nacional de Águas e Saneamento Básico (ANA).

2.2.2 Obtenção dos dados da concentração superficial de sedimentos (CSS)

Os dados observados da concentração superficial de sedimentos (CSS), utilizados como variável resposta, foram valores médios da seção transversal medidos em sete estações sedimentométricas (Figura 2.1) pertencentes à rede hidrometeorológica da Agência Nacional de Águas e Saneamento Básico (ANA). Os dados foram obtidos por meio do portal HidroWeb do Sistema Nacional de Informações sobre Recursos Hídricos (SNIRH) (HIDROWEB, 2022).

A Tabela 2.1 apresenta o código das estações sedimentométricas utilizadas, sua localização geográfica, descrição, período de observação e número de observações disponíveis.

Tabela 2.1. Estações sedimentométricas da ANA utilizadas na modelagem da concentração superficial de sedimentos (CSS)

Long.	Lat.	Código	Nome da estação	Período	Nº de observações
-43,088	-20,670	56075000	Porto Firme	1998-2019	57
-42,903	-20,384	56110005	Ponte Nova Jusante	1982-2019	88
-42,674	-19,994	56425000	Fazenda Cachoeira das Antas	1998-2019	42
-42,476	-19,777	56539000	Cachoeira dos Óculos Montante	1999-2019	43
-41,642	-18,971	56920000	Tumiritinga	1974-2019	71
-41,003	-19,509	56850000	Governador Valadares	1998-2019	29
-40,630	-19,533	56994500	Colatina	1977-2019	78

Realizou-se a individualização da área de drenagem de cada uma das sete estações sedimentométricas a partir do Modelo Digital de Elevação Hidrograficamente Condicionado (MDEHC), com 30 m de resolução espacial, por meio da ferramenta *r.watershed* do software *Geographic Resources Analysis Support System* (GRASS), integrado ao software QGIS versão 3.16.9 (QGIS CORE TEAM, 2021). A delimitação dessas áreas é fundamental para a obtenção das variáveis espaciais que tenham influência na CSS obtida em cada estação sedimentométrica. Na Figura 2.2 é possível observar a individualização da área de drenagem de cada uma das sete estações sedimentométricas utilizadas.

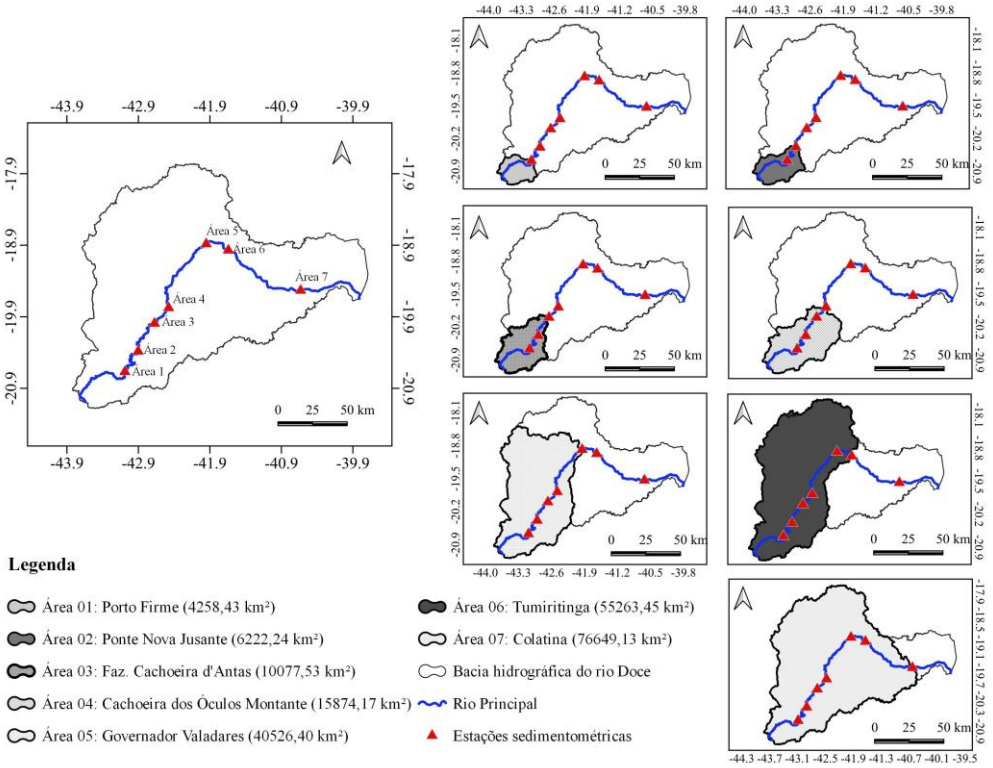


Figura 2.2. Individualização da área de drenagem a montante de cada uma das sete estações sedimentométricas utilizadas no estudo.

Para aumentar a quantidade de dados disponíveis no treinamento dos modelos de aprendizado de máquina, as estações sedimentométricas foram agregadas, formando duas bases de dados: uma definida como região alto rio Doce, com 230 dados, que incluiu as estações 56075000 (área 01), 56110005 (área 02), 56425000 (área 03) e 56539000 (área 04); e outra para a região definida como baixo rio Doce, com as estações 56920000 (área 05), 56850000 (área 07) e 56994500 (área 08), com 178 dados observados da CSS.

2.2.3 Obtenção das variáveis utilizadas na predição da concentração superficial de sedimentos

As variáveis preditivas foram obtidas para a área de drenagem a montante de cada uma das sete estações sedimentométricas localizadas na calha do rio Doce e utilizadas neste estudo (Figura 2.2). Foram empregadas como variáveis preditoras os seguintes dados: declividade, pedologia, uso e cobertura da terra, precipitação, vazão fluvial, velocidade fluvial, evapotranspiração real, escoamento superficial, umidade do solo, temperatura e *normalized difference vegetation index* (NDVI).

Dados de declividade foram extraídos a partir do Modelo Digital de Elevação Hidrograficamente Condicionado (MDEHC), gerado por meio do Modelo Digital de Elevação (MDE) do projeto *Shuttle Radar Topography Mission* (SRTM), com resolução espacial de 30 m (FARR et al., 2007). Os cálculos de declividade foram feitos com auxílio do software QGIS versão 3.16.9. Foram obtidas as seguintes informações de declividade: declividade média de cada área de drenagem (Figura 2.2) em porcentagem, declividade do canal principal (m/m) e as informações de declividade de acordo com a classificação da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA).

Para a definição da declividade do canal principal foi utilizado método da declividade equivalente constante (S), média harmônica ponderada da raiz quadrada da declividade, dada pela Equação 2.1 (MELLO; SILVA, 2013),

$$S = \left(\frac{\sum_{i=1}^{n} Li}{\sum_{i=1}^{n} \left(\frac{Li}{Di} \right)} \right)^2 \qquad (2.1)$$

em que, S é a declividade equivalente constante, em $m.m^{-1}$, Li é o comprimento do trecho i, em m, e Di é a declividade em cada trecho i, em $m.m^{-1}$

A classificação das declividades pela EMBRAPA se divide nas categorias apresentadas na Tabela 2.2.

Tabela 2.2. Classes de declividade utilizadas

Declividade (%)	Relevo
0 - 3	Plano
3 - 8	Suave-ondulado
8 - 20	Ondulado
20 - 45	Forte-ondulado
45 - 75	Montanhoso
> 75	Forte-montanhoso

Fonte: EMBRAPA (2018)

Realizou-se a classificação da declividade por meio da ferramenta *r.classify* do software GRASS, integrado ao software QGIS versão 3.16.9. Foram extraídos os valores de área (km²) que cada categoria de declividade representa nas áreas de drenagem da respectiva estação sedimentométrica.

Os dados de pedologia foram extraídos do arquivo vetorial de mapas de solos, na escala de 1:250.000, elaborado Instituto Brasileiro de Pesquisa e Estatística (IBGE) (IBGE, 2019). Foram obtidos os valores de área (km²) dos tipos de solos predominantes nas áreas de drenagem de cada uma das estações sedimentométricas, os quais consistem em: Nitossolo Vermelho Eutróficos (NVe), Luvissolos Crômicos Órticos (TCo), Argissolo Vermelho Eutrófico (PVe), Cambissolos Háblicos Tb Distróficos (CXbd), Neossolos Flúvicos Ta Eutróficos (RYve), Latossolos Vermelhos Perférricos (LVj).

Os usos e cobertura da terra foram obtidos do Projeto MapBiomas na sua versão 5, que disponibiliza mapas anuais da dinâmica da cobertura da superfície para todo o Brasil com resolução espacial de 30 m, para o período 1985 a 2019, com base na classificação de imagens da série de satélites Landsat (MAPBIOMAS, 2022). Estes dados estão disponíveis na plataforma computacional de processamento de dados em nuvem *Google Earth Engine* (GEE) (GEE, 2017) e foram extraídos para a região de estudo a partir desta ferramenta.

Foram calculadas as áreas (km²) das seguintes classes de uso e cobertura da terra: floresta nativa, floresta plantada, formação natural não florestal, pastagens, agricultura, mosaico agricultura e pastagem, mineração, área não vegetada e afloramento rochoso. Esse procedimento foi feito usando o software QGIS versão 3.16.9, por meio do plugin *Land Cover Statistics* (LecoS), para cada área de drenagem das estações sedimentométricas utilizadas.

Os mapas de uso e cobertura da terra disponibilizados pelo projeto MapBiomas na coleção 5 apresentam 30 classes. As classes de maior interesse foram reclassificadas, e aquelas com características espectrais semelhantes foram agregadas. A Tabela 2.3 apresenta a reclassificação das classes de uso e cobertura da terra que foram empregadas neste estudo, utilizando ferramenta *r.classify*.

Tabela 2.3. Reclassificação das classes de uso e cobertura da terra do projeto MapBiomas utilizadas no estudo

Classes MapBiomas	Identidade	Classes reclassificadas
Floresta natural	2	Floresta nativa
Formação Florestal	3	
Formação Savânica	4	
Floresta Plantada	9	Floresta Plantada
Campo Alagado e Áreas Pantanosas	11	Formação não florestal
Formação Campestre	12	
Outras formações não florestais	13	
Apicum	32	Formação não florestal
Pastagem	15	Pastagem
Lavoura Temporária	19	Agricultura
Soja	39	
Cana	20	
Outras Lavouras Temporárias	41	
Lavoura Perene	36	
Mosaico de Agricultura e Pastagem[3]	21	Mosaico Agricultura e Pastagem
Praia e Duna	23	Área não vegetada
Infraestrutura Urbana	24	
Outras Áreas não Vegetadas	25	
Afloramento Rochoso	29	Afloramento Rochoso
Mineração	30	Mineração

[3]mosaico de agricultura e pastagem são áreas em que não foi possível a distinção entre essas duas classes.

Em relação as informações de precipitação pluvial, foram utilizadas duas bases de dados. A primeira consistiu em dados observados de precipitação das estações pluviométricas pertencentes à rede da ANA e disponibilizados no portal HidroWeb (HIDROWEB, 2022). A localização geográfica, o código e o período de disponibilidade de dados dessas estações podem ser verificados no Apêndice I.

Foram utilizados dados de 79 estações pluviométricas localizadas na área de drenagem da bacia hidrográfica do rio Doce (Figura 2.1), tendo sido obtido o valor médio diário de

precipitação utilizando o interpolador determinístico *Inverse Distance Weighted (IDW),* com expoente dois, por meio do software R versão 4.1.1 (R CORE TEAM, 2021).

As precipitações interpoladas consistiram naquelas correspondentes à data da coleta da CSS e, ainda, nas chuvas defasadas no tempo em até cinco dias antes da coleta, ou seja, as chuvas ocorridas em 24h, 48h, 72h, 96h e 120h anteriormente à coleta dos sedimentos. Foram utilizadas como variáveis explicativas dos modelos de aprendizado de máquina, além das precipitações médias diárias, também as precipitações médias acumuladas no período de 24h, 48h, 72h, 96h e 120h antes da data da coleta dos sedimentos. Os valores médios de precipitação na área de drenagem de cada uma das sete estações sedimentométricas foram extraídos por meio do software QGIS versão 3.16.9.

A segunda base de dados foi a de precipitações diárias estimadas pelo projeto *Climate Hazards Group InfraRed Precipitation With Station Data* (CHIRPS) (CHIRPS, 2022), o qual é um produto de reanálise que combina dados de precipitação medidos em estações pluviométricas e dados de sensoriamento remoto orbital. Os dados estão disponíveis a partir de 1981, em escala global, com resolução espacial de 0,05º (≈ 5 km) e disponíveis em escala diária (FUNK et al., 2015).

A aquisição e processamento dos dados do CHIRPS foi feito por meio da plataforma GEE, em que se filtrou as imagens de precipitação para o período de interesse e limites espaciais referentes a cada uma das áreas de drenagem das sete estações sedimentométricas, e exportou-se as séries temporais com valores médios diários. Essas séries temporais corresponderam às precipitações na data da coleta e defasadas no tempo em até cinco dias antes da coleta. Também foram obtidos os valores de precipitação acumulados em 24h, 48h, 72h, 96h e 120h anteriores à data da coleta.

Os dados de velocidade e vazão fluvial foram obtidos por meio do portal HidroWeb (HIDROWEB, 2022). O código das estações utilizadas são os mesmos das estações sedimentométricas apresentados na Tabela 2.1. No Brasil, as estações sedimentométricas são instaladas em locais em que já ocorrem as medições de vazão, de modo a viabilizar a obtenção das curvas de retenção de sedimentos (SRC) (CARVALHO et al., 2000).

A velocidade fluvial foi obtida para a mesma data da coleta de CSS em cada estação sedimentométrica utilizada. Os dados de vazão corresponderam à data da coleta da CSS e também defasagens em até cinco dias antes da coleta. É esperado que exista boa correlação entre os dados da CSS e as vazões defasadas no tempo, pois é observado um atraso entre o aumento da vazão e o subsequente aumento nos sedimentos em suspensão, devido ao fenômeno

da histerese (CAO et al., 2021; HADDADCHI; HICKS, 2021). Também foram utilizadas as vazões médias, mínimas e máximas ocorridas no período de 30 dias antes da data da coleta.

Os dados de Evapotranspiração real (ET), escoamento superficial (ro) umidade do solo (U) e temperatura (T) (média, máxima e mínima) foram obtidos para o mês em que ocorreram as coletas de CSS. Essas variáveis foram extraídas do banco de dados TerraClimate, que fornece informações mensais do clima e do balanço de água na superfície terrestre em escala global, disponíveis a partir de 1958 e com resolução espacial de 0,04º (≈ 4 km) (ABATZOGLOU et al., 2018). A aquisição das séries temporais do TerraClimate foi feita por meio da plataforma GEE, em que se filtrou os dados matriciais de ET, ro, U e T para o período de interesse e limites das áreas de drenagem das estações sedimentométricas utilizadas, tendo sido exportado os dados mensais referentes ao mês da coleta de CSS.

As séries temporais de NDVI utilizadas como variáveis explicativas foram obtidas por meio do *National Oceanic and Atmospheric Administration* (NOAA) *Climate Data Record* (CDR), a qual fornece dados mensais de NDVI em escala global, com resolução espacial de 0,05º (≈ 5 km) a partir de 1981 (PEDELTY et al., 2007). A aquisição dos dados foi feita por meio da plataforma GEE, em que se filtrou os *rasters* de NDVI para o período correspondente ao mês da coleta da CSS e limites espaciais das áreas de drenagem de cada uma das estações sedimentométricas utilizadas. As series temporais mensais foram exportadas no formato CSV.

Na Tabela 2.4 são apresentadas as abreviações das 62 variáveis preditoras utilizadas nos algoritmos de seleção de variáveis para a sua identificação, descrição e unidades de medidas.

Tabela 2.4. Siglas utilizadas para a variáveis preditoras, descrição e unidades de medidas

Sigla da variável	Descrição	Unidade
Slope_S3	Declividade do canal (SRTM)	$m.m^{-1}$
plano_0_3	Área de declividade de 0 a 3%	km²
suave_3_8	Área de declividade de 3 a 8%	km²
ondulado_8_20	Área de declividade de 8 a 20%	km²
forte_ondulado_20_45	Área de declividade de 20 a 45%	km²
montanhoso_45_75	Área de declividade de 45 a 75%	km²
escarpado_75	Área de declividade maior que 75%	km²
slope_mean	Declividade média da área de drenagem	km²
luvissolo	Área de solos ocupadas por luvissolos	km²
latossolo	Área de solos ocupadas por latossolos	km²
argissolo	Área de solos ocupadas por argissolos	km²
nitossolo	Área de solos ocupadas por nitossolos	km²
cambissolo	Área de solos ocupadas por cambissolos	km²
neossolo	Área de solos ocupadas por neossolos	km²
floresta_nativa	Área de uso da terra ocupada por floresta nativa	km²
floresta_plantada	Área de uso da terra ocupada por floresta plantada	km²
formacao_nao_florestal	Área de uso da terra ocupada por formação não florestal	km²
pastagem	Área de uso da terra ocupada por pastagem	km²
agricultura	Área de uso da terra ocupada por agricultura	km²
mosaico_agri_past	Área de uso da terra ocupada por mosaico de agricultura e pastagem	km²
area_nao_vegetada	Área de uso da terra ocupada por área não vegetada	km²
afloramento_rochoso	Área de uso da terra ocupada por afloramento rochoso	km²
mineracao	Área de uso da terra ocupada por mineração	km²
p_0	Precipitação média diária do dia da coleta de sedimentos (pluviômetro)	mm
p_1	Precipitação média diária de um dia antes coleta de sedimentos (pluviômetro)	mm
p_2	Precipitação média diária de dois dias antes coleta de sedimentos (pluviômetro)	mm

Continua...

Tabela 2.4 Continuação...

p_3	Precipitação média diária de três dias antes coleta de sedimentos (pluviômetro)	mm
p_4	Precipitação média diária de quatro dias antes coleta de sedimentos (pluviômetro)	mm
p_5	Precipitação média diária de cinco dias antes coleta de sedimentos (pluviômetro)	mm
p_acc_2	Precipitação média diária acumulada p_0 a p_1 antes coleta de sedimentos (pluviômetro)	mm
p_acc_3	Precipitação média diária acumulada p_0 a p_2 antes coleta de sedimentos (pluviômetro)	mm
p_acc_4	Precipitação média diária acumulada p_0 a p_3 antes coleta de sedimentos (pluviômetro)	mm
p_acc_5	Precipitação média diária acumulada p_0 a p_4 antes coleta de sedimentos (pluviômetro)	mm
p_acc_6	Precipitação média diária acumulada p_0 a p_5 antes coleta de sedimentos (pluviômetro)	mm
p_5_ch	Precipitação média diária de cinco dias antes coleta de sedimentos (CHIRPS)	mm
p_4_ch	Precipitação média diária de quatro dias antes coleta de sedimentos (CHIRPS)	mm
p_3_ch	Precipitação média diária de três dias antes coleta de sedimentos (CHIRPS)	mm
p_2_ch	Precipitação média diária de dois dias antes coleta de sedimentos (CHIRPS)	mm
p_1_ch	Precipitação média diária de um dia antes coleta de sedimentos (CHIRPS)	mm
p_0_ch	Precipitação média diária do dia da coleta de sedimentos (CHIRPS)	mm
p_acc_2_ch	Precipitação média diária acumulada p_0_ch a p_1_ch antes coleta de sedimentos (CHIRPS)	mm
p_acc_3_ch	Precipitação média diária acumulada p_0_ch a p_2_ch antes coleta de sedimentos (CHIRPS)	mm
p_acc_4_ch	Precipitação média diária acumulada p_0_ch a p_3_ch antes coleta de sedimentos (CHIRPS)	mm
p_acc_5_ch	Precipitação média diária acumulada p_0_ch a p_4_ch antes coleta de sedimentos (CHIRPS)	mm
p_acc_6_ch	Precipitação média diária acumulada p_0_ch a p_5_ch antes coleta de sedimentos (CHIRPS)	mm
vel_water	Velocidade da água na data da coleta (ANA)	$m.s^{-1}$
q_5	Vazão diária defasa em 5 dias anteriores a coleta de sedimentos	$m^3.s^{-1}$
q_4	Vazão diária defasa em 5 dias anteriores a coleta de sedimentos	$m^3.s^{-1}$
q_3	Vazão diária defasa em 3 dias anteriores a coleta de sedimentos	$m^3.s^{-1}$
q_2	Vazão diária defasa em 2 dias anteriores a coleta de sedimentos	$m^3.s^{-1}$
q_1	Vazão diária defasa em 1 dia anterior a coleta de sedimentos	$m^3.s^{-1}$
q_0	Vazão diária do dia da coleta de sedimentos	$m^3.s^{-1}$
q_max_month	Vazão máxima de 30 dias anteriores a coleta de sedimentos	$m^3.s^{-1}$

Continua...

Tabela 2.4 Continuação...

q_min_month	Vazão mínima de 30 dias anteriores a coleta de sedimentos	$m^3.s^{-1}$
q_med_month	Vazão média de 30 dias anteriores a coleta de sedimentos	$m^3.s^{-1}$
ET	Evapotranspiração real do mês da coleta	mm
ro	Escoamento superficial do mês da coleta	mm
soil	Umidade do solo do mês da coleta	mm
tmin	Temperatura mínima do mês da coleta	°C
tmax	Temperatura máxima do mês da coleta	°C
tmed	Temperatura média do mês da coleta	°C
NDVI_NOAA	NDVI médio do mês da coleta (NOAA)	Adimensional

2.2.4 Pré-processamento da base de dados e métodos de seleção de variáveis

O pré-processamento dos dados consistiu nas etapas de identificação e remoção de outliers, padronização dos dados, preenchimento de dados perdidos e seleção do conjunto ótimo de variáveis a serem utilizadas nos modelos de aprendizado de máquina. Esses procedimentos são necessários para melhorar a capacidade preditiva dos modelos e reduzir o custo computacional no processamento dos dados.

A identificação e remoção de outliers é bastante complexa, pois apesar desses valores dificultarem a modelagem hidrológica, podem representar eventos atípicos que não devem ser excluídos sem analisar o comportamento hidrológico no período (RITTER; MUÑOZ-CARPENA, 2013). Desta forma, a remoção de *outliers* foi feita de forma manual, comparando-se os valores da CSS com dados observados da vazão fluvial na data da coleta e em dias anteriores, plotando-se os valores em um gráfico.

Como as variáveis utilizadas estão em diferentes escalas, foi feito a padronização dos dados utilizando o método z-score que é a relação entre o valor observado menos a média, dividido pelo desvio padrão. Este método é indicado quando o conjunto de variáveis apresentam unidades e dispersões bastante heterogêneas (HAIR et al., 2009). Destaca-se, ainda, que alguns algoritmos de aprendizado de máquina não trabalham bem com dados em diferentes escalas (JAMES et al., 2021).

No conjunto de variáveis preditoras foi considerado apenas aquelas que apresentaram até 10% de falhas. Para o preenchimento das falhas foi empregado o método do *K-Nearest Neighbors* (KNN), utilizando a função para imputação de dados faltantes do pacote Caret (KUHN, 2021) do software R versão 4.1.1, a qual considera o valor médio entre os valores de maior ocorrência dentro do conjunto de dados (JAMES et al., 2021).

Com o intuito de diminuir o número de variáveis preditoras, eliminando aquelas redundantes, com pouca relevância, e selecionar o conjunto ótimo de variáveis, foram utilizados os métodos supervisionados de seleção de variáveis Boruta e *Recursive Featuare Elimination (RFE)* com a função *random forest* (RF).

O método de seleção de variáveis Boruta utiliza o z-score do algoritmo *random forest* como indicativo da importância da variável. Este representa a perda de acurácia média em cada árvore dividida pelo seu desvio padrão. O algoritmo compara a capacidade preditiva de cada variável explicativa com suas versões randomizadas (s*hadow shuffled feature*), ou seja, compara a relação y_1 e x_1 com valores aleatórios do conjunto de dados x_i, repetindo-se esse processo para todas as variáveis preditoras. É selecionado o valor máximo de z-score obtido

para as variáveis randomizadas (MZSA). Se o valor de z-score das variável originais forem significantemente superior ao MZSA, essas variáveis são consideradas importantes para a modelagem (KURSA; JANKOWSKI; RUDNICKI, 2010). O algoritmo Boruta foi implementado a partir do pacote Boruta (KURSA e RUDNICKI, 2010) por meio do software R versão 4.1.1.

A seleção do conjunto ótimo de variáveis por meio do RFE-RF é feita com base no erro de predição do algoritmo *random forest*. Para cada árvore de decisão há um conjunto amostral (*out of bag*) que não foi utilizado para o treinamento do modelo, o qual é empregado para calcular o erro de predição não viesado. Com RFE-RF o erro de predição de cada variável é comparado ao erro de predição de suas versões randomizadas (*shuffled feature*), similar ao procedimento utilizado pelo algoritmo Boruta. Variáveis irrelevantes para o modelo não apresentam diferenças significativas nos erros de predição comparados com suas versões randomizadas (GRANITTO et al., 2006). O algoritmo RFE-RF foi implementado por meio do pacote Caret utilizando o software R versão 4.1.1.

2.2.5 Modelos de aprendizado de máquina utilizados para a predição da CSS

Os modelos implementados baseiam-se em algoritmos de aprendizado de máquina, os quais tem demostrando bons resultados no monitoramento dos sedimentos em suspensão. Foram utilizados os algoritmos *Random Forest (RF)*, *Cubis*t, *Support Vector Machines* (SVMs), *Extreme Gradient Boosting Machine* (XGBoost) e regressão *Least Absolute Shrinkage and Selection Operator* (LASSO). Os modelos foram implementados utilizando o pacote Caret do software R versão 4.1.1.

O algoritmo RF é baseado na utilização de múltiplas árvores de decisão, em que cada uma é construída de forma independente por meio de um subconjunto amostral (*bootstrapped*), obtido aleatoriamente a partir dos dados originais. Esse processo é repetido várias vezes, resultando em diferentes combinações das árvores, o que torna o RF mais efetivo do que as árvores de decisões individuais. Os resultados dos diversos modelos ajustados em cada árvore são agregados no processo decisório para construção do modelo final, conhecido como *bagging* (TANIGUCHI; SATO; SHIRAKAWA, 2018).

O processo decisório, que inclui os melhores modelos na composição do modelo final, baseia-se no erro de predição não viesado obtido do subconjunto amostral *out- of- bag* (OOB), que é uma fração dos dados selecionados no processo de *bootstrapping* não utilizados no treinamento dos modelos (GRANITTO et al., 2006).

O algoritmo Cubist é uma extensão do modelo de árvores de Quinlan M5 (QUINLAN, 1992), o qual consiste na combinação de modelos de regressão linear e árvores de decisão. Inicialmente este particiona as variáveis respostas em subconjuntos com características semelhantes (regras hierárquicas). Em seguida, o algoritmo aplica regressão linear a esses subconjuntos. Desta forma, no modelo Cubist o valor médio da predição que é utilizado nos modelos de árvore de decisão é substituído pela regressão linear em cada folha.

O Cubist ainda utiliza duas funções nas predições. A primeira consiste na utilização da função *neighbor*, que aplica o algoritmo *nearest neighbor* em cada folha e o combina com a predição obtida no subconjunto. A segunda função consiste na *committee*, a qual é similar ao *boosting*, ou seja, os modelos ajustados subsequentes consideram os erros de predição ocorridos nos ajustes anteriores (BUTLER; O'ROURKE; HILLIER, 2018).

Os SVMs transformam os dados observados, que estão em uma única dimensão (1D), e os compara com o seu valor transformado no hiperplano. Essa transformação auxilia na identificação de padrões que em dimensões do espaço menores seriam de difícil detecção. Em seguida ajusta vetores que separam e identificação padrões nos dados (*Support Vectors*). Na separação dos hiperplanos considera-se uma margem flexível, ajustada por meio de validação cruzada, que permite adição de viés ao modelo, objetivando a redução da variância (HEARST et al., 1998).

O custo computacional da transformação dos dados é elevado, portanto o SVMs trabalha com funções Kernel, que por meio do truque de Kernel compara os dados observados em dimensões mais altas sem precisar transformá-los. Neste estudo foi utilizado o SVMs com Kernel da função de base radial (RBF), que permite a comparação dos dados em infinitas dimensões através da extensão das séries de Taylor. Além disso, a RBF é amplamente utilizada com SVMs por se assemelharem ao algoritmo KNN, considerando o peso do vizinho mais próximo entre os dados na separação dos hiperplanos (NOROUZI; DANESHFARAZ; GHADERI, 2019).

O algoritmo XGBoost baseia-se no ajuste de múltiplas árvores de decisão para minimizar o resíduo entre os dados observados e preditos, proposto por Chen e Guestrin (2016). Em linhas gerais, o modelo inicia as árvores adicionando todos os resíduos em uma única folha, e calcula o grau de similaridade (*similarity score*) entre os resíduos. Após esse processo, inicia-se a divisão dos resíduos em diferentes clusters (folhas) e novamente calcula-se o grau de similaridade.

Para avaliar se houve vantagem na divisão dos resíduos em diferentes clusters, calcula-se o ganho (*gain*), o qual é uma relação entre o grau de similaridade entre as folhas menos

aquele obtido contendo todos os resíduos (raiz). Os maiores valores de ganho indicam a melhor ramificação para a divisão dos resíduos até um critério de parada. Isso é fundamental pois o vetor de saída de cada folha utilizado na função de custo do modelo é a relação entre a somatória dos resíduos em cada folha dividido pelo total de resíduos adicionado pelo termo de regularização (λ) (NI et al., 2020).

Diante disso, o XGBoost visa minimizar a função objetivo (Obj^t) expressa na Equação 2.2.

$$Obj^{(t)} = \sum_{k=1}^{n} l\left(y_i, y_i\right) + \sum_{k=1}^{t} \Omega\left(f_i\right) \tag{2.2}$$

em que l é a função de custo, n é o número de observações, $\hat{y}_i$ é o valor estimado, y_i é o valor observado e Ω é um termo de regularização do modelo XGBoost obtido por meio da Equação 2.3

$$\Omega\left(f\right) = \gamma T + \frac{1}{2}\lambda\left\|\omega\right\|^2 \tag{2.3}$$

em que ω é o vetor de saída de cada folha da árvore de decisão, λ é um termo de regularização similar a penalização do tipo L_2 da regressão *Ridge*, utilizado para evitar *overfitting*, T é o número de folhas na árvore de decisão, γ é um termo de complexidade de cada folha.

A regressão LASSO penaliza a inserção de novas varáveis que apresentam pouca capacidade preditiva por meio do processo de regularização (JAMES et al., 2021). A regressão linear ajusta os coeficientes de regressão β_i de modo a minimizar a função de custo, soma residual dos quadrados (SRQ). Já a regressão LASSO penaliza os coeficientes de regressão (β_j) acrescentado um novo termo na SRQ, conforme a Equação 2.4. Esse termo é a penalização do tipo L1, a qual tem o efeito de zerar os coeficientes quando o valor de λ for suficientemente grande, pois à medida que este aumenta, menor fica a inclinação da reta, sendo particularmente útil quando há um grande número de variáveis passíveis de serem removidas (JAMES et al., 2021).

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \quad (2.4)$$

em que y_i é o valor observado, β_j e β_0 são os coeficientes de regressão, x_{ij} é a variável explicativa, e λ é um hiperparâmetro da equação que vária de 0 a $+\infty$, obtido utilizando validação cruzada para identificar o valor de λ que resulta na menor variância.

2.2.6. Método de validação cruzada utilizado e métricas para avaliação dos modelos

Foi selecionado o método de validação cruzada *leave-one-out* (LOOCV), para o treinamento e teste dos modelos de aprendizado de máquina utilizados, implementado no software R versão 4.1.1, por meio do pacote Caret. A LOOCV é um caso especial de validação cruzada *k-fold*, o qual também envolve a divisão da base de dados em duas partes para treinamento e teste, no entanto, ao invés de criar dois subconjuntos, apenas uma observação (x_1, y_1) é usada para o teste e o restante das observações $\{(x_2, y_2),..., (x_n, y_n)\}$ são empregadas no treinamento. O modelo de predição utilizado é ajustado em n -1 observações, e uma predição $\hat{y}_1$ é feita para a observação x_1 que não entrou no conjunto de treinamento (JAMES et al., 2021). Esse procedimento é repetido até que todo o conjunto de dados tenha sido utilizado para treinamento e teste.

A vantagem desse método consiste na redução do viés em função da repetição do ajuste do modelo em n-1 vezes. O modelo preditivo gerado é mais estável, pois a divisão dos subconjuntos de treinamento não é feita de forma aleatória, como é o caso dos demais tipos de *k-fold,* em que os resultados podem ser diferentes dependendo dos dados selecionados. Entretanto, para um grande conjunto de dados o método LOOCV demanda computadores com grande capacidade de processamento (JAMES et al., 2021).

As métricas utilizadas para avaliação dos modelos em relação a concordância dos dados preditos com os observados da CSS foram: Erro Médio Absoluto (MAE), Raiz do Erro Médio Quadrático (RMSE), Porcentagem do Viés (PBIAS), coeficiente de Nash–Sutcliffe (NSE), índice de concordância de Willmot (d), coeficiente de determinação (R^2), coeficiente de Kling-Gupta (KGE) e índice de desempenho (c). As métricas de avaliação dos modelos foram obtidas utilizando o software R versão 4.1.1, por meio do pacote hydroGOF (ZAMBRANO-BIGIARINI, 2020).

O MAE foi obtido por meio da Equação 2.5 na mesma unidade da variável analisada. Quanto menor for o valor de MAE obtido na performance do modelo, mais acurado. O ideal é a obtenção de valores próximo a zero, além disso essa métrica serve como indicativo da

presença de outlier nos dados utilizados, em situações em que o RMSE for superior ao MAE (RITTER; MUÑOZ-CARPENA, 2013).

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|O_i - P_i\right| \tag{2.5}$$

em N é o número de elementos da amostra, P_i é o valor dos dados preditos, O_i é o valor dos dados observados.

A RMSE é a medida mais comumente utilizada para aferir a qualidade do ajuste dos modelos, caracterizada por ser uma medida análoga ao desvio padrão e valores do erro nas mesmas dimensões da variável analisada (HALLAK; PEREIRA FILHO, 2011), obtida pela Equação 2.6. Quanto mais próximos de zero, melhor é a capacidade preditiva do modelo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - O_i\right)^2} \tag{2.6}$$

em que n é o número de elementos da amostra, P_i é o valor dos dados estimados, O_i é o valor dos dados observados.

A Pbias indica a diferença entre os valores preditos e observados. Valores baixos de Pbias indicam boa performance do modelo, sendo que zero é o valor ideal. Valores positivos indicam superestimativas, enquanto que valores negativos subestimativas dos modelos (LI et al., 2009). A Equação 2.7 demostra a fórmula de obtenção dessa métrica.

$$Pbias = \left[\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)100}{\sum_{i=1}^{n}O_i}\right] \tag{2.7}$$

em que P_i é o valor dos dados estimados e O_i é o valor dos dados observados.

O NSE é um índice de eficiência adimensional variam de -∞ a 1, onde NSE igual a 1 indica ajuste perfeito dos dados, enquanto que valores menores que 0, sugerem que a média dos dados observados é melhor que o modelo ajustado (RITTER; MUÑOZ-CARPENA, 2013). Este pode ser obtido com base na Equação 2.8.

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(O_i - P_i\right)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}_i\right)^2} \tag{2.8}$$

em que n é o número de elementos da amostra, P_i é o valor dos dados estimados, O_i é o valor dos dados observados, $\bar{O}_i$ é a média dos valores observados.

O NSE é amplamente utilizado em recursos hídricos devido a sua flexibilidade para trabalhar com diversos modelos matemáticos. Este índice apresenta maior sensibilidade ao viés nas predições dos modelos e ao efeito de *outliers* presentes nas séries de dados (ALTHOFF; RODRIGUES, 2021).

O índice d proposto por Wilmott é empregado para identificar o grau de concordância entre o valor observado e sua estimativa, e pode ser obtido por meio da Equação 2.9.

$$d = 1 - \left[\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{\sum_{i=1}^{n}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2}\right] \tag{2.9}$$

em que n é o número de elementos da amostra, $\bar{O}$ é a média dos valores observados, Oi representa os dados observados e Pi os valores estimados.

Este índice é de simples interpretação, sendo que seus valores variam de 0 a 1, sendo 1 o ajuste perfeito do modelo utilizado. No entanto, este apresenta a desvantagem quanto ao uso das diferenças quadráticas em seu algoritmo, o qual pode resultar em valores altos (bons ajustes) mesmo se a capacidade preditiva do modelo for ruim (PEREIRA et al., 2018).

O R^2 também é utilizado para verificar o ajuste dos modelos em relação a sua capacidade de predizer um fenômeno. Pode ser expresso por meio da Equação 2.10.

$$R^2 = \frac{\sum\left(\hat{P}_i - \bar{O}\right)^2}{\sum\left(O_i - \bar{O}\right)^2} \tag{2.10}$$

em que $\bar{O}$ é a média dos valores observados, O_i representa os dados observados e Pi os valores estimados.

A interpretação dos resultados obtidos do R^2 são similares aos do coeficiente de d, ou seja, variando de 0 a 1, os valores ideais são aqueles próximos da unidade. Uma das principais

desvantagens do R^2 é a quantificação apenas de dispersão (variação) dos dados se for considerado isoladamente. Um modelo que sub ou superestima sistematicamente a variável observada ao longo do tempo poderá ter valores de R^2 próximos de 1, mesmo se todas as estimativas estiverem erradas (NAGUETTINI; PINTO, 2007).

O KGE envolve a avaliação de três componentes entre os dados preditos e observados, os quais são: correlação, viés e medidas de variabilidade (GUPTA et al., 2009), obtido por meio da Equação 2.11. Os valores podem variar de -∞ a 1, sendo 1 o valor do ajuste perfeito dos dados pelo modelo.

$$KGE = 1 - ED = 1 - \sqrt{\left[s_r\left(r-1\right)\right]^2 + \left[s_\alpha\left(\alpha-1\right)\right]^2 + \left[s_\beta\left(\beta-1\right)\right]^2} \quad (2.11)$$

em que ED é a distância euclidiana; r é o coeficiente de correlação entre O_i e P_i; α é a razão entre o desvio padrão dos dados preditos com o desvio padrão dos dados observados; β é a razão entre a média dos valores preditos pela média dos valores observados; s_r, s_α e s_β são fatores de escala.

O índice de eficiência (c) foi proposto por Camargo e Sentelhas (1997) e permite a classificação dos modelos ajustados para dar suporte na escolha daquele que apresentou melhor desempenho. Esse índice é o resultado do produto entre o coeficiente de Willmott (d) e o coeficiente de correlação de Pearson (r), sendo que valores $c \leq 0,4$ indicam que os modelos são classificados como péssimos; entre 0,41 a 0,5: mal; entre 0,51 a 0,60: sofrível; entre 0,61 a 0,65: mediano; entre 0,66 a 0,75: bom; entre 0,76 a 0,85, muito bom; e c > que 0,85 indica que os modelos são classificados como ótimos.

2.3. RESULTADOS E DISCUSSÃO

2.3.1. Dados de concentração superficial de sedimentos (CSS)

Na Tabela 2.5 é possível analisar as estatísticas descritivas dos dados de CSS de cada uma das sete estações sedimentométricas utilizados neste trabalho.

Tabela 2.5. Estatística descritiva dos dados de CSS em cada estação sedimentométrica

	Alto rio Doce				Baixo rio Doce		
Estatística	**56075000**	**56110005**	**56425000**	**56539000**	**5685000**	**56920000**	**56994500**
Mínimo	2,10	2,30	2,20	12,04	3,53	2,80	1,29
Máximo	240,22	350,04	179,10	1764,00	240,71	476,22	503,53
Range	238,12	347,74	176,90	1751,96	237,18	473,42	502,24
Mediana	19,92	18,53	48,50	45,11	24,81	52,56	42,70
Média	44,18	46,94	52,29	196,01	53,07	108,84	77,38
Variância	3393,47	4106,00	1450,17	169151,28	4807,15	17466,31	9354,05
Desvio Padrão	58,25	64,08	38,08	411,28	69,33	132,16	96,72
Coef. de Variação	1,32	1,37	0,73	2,10	1,31	1,21	1,25

Analisando a Tabela 2.5, verifica-se que o maior desvio padrão foi observado na estação 56539000, localizada na região alto rio Doce. Isso ocorreu devido a três medições da CSS que estão muito acima da média, com valores de 1764 mg/L, 1692 mg/L e 1288 mg/L, ocorridos em 24/01/2016, 10/12/2015 e 07/12/2000, respectivamente.

Entretanto, nesses períodos foram observadas vazões diárias na data da coleta dos sedimentos com valores próximos à vazão máxima observada nesta estação, ou seja, 569,98 m³/s (24/01/2016), 438,39 m³/s (10/12/2015) e 518,47 m³/s (07/12/2000). O total precipitado também foi bastante elevado, com valores médios acumulados de seis dias antes das datas das coletas de 82,49 mm, 87,82 mm e 61,65 mm, respectivamente.

Além disso, os valores elevados de sedimentos em suspensão estão relacionados ao rompimento da barragem de fundão ocorrido em 05 de novembro de 2015, que liberou grande quantidade rejeitos de mineração na bacia hidrográfica do rio Doce (AIRES et al., 2018). Essa estação está localizada na região próxima ao deságue dos rejeitos na calha do rio Doce. Como as datas de medições são diferentes para cada estação e, em decorrência das diferenças na dinâmica da sedimentologia fluvial, é difícil observar valores tão elevados de CSS nas demais estações. A velocidade fluvial e os locais de represamento influenciam diretamente a quantidade de partículas em suspensão, pois afetam a capacidade de transporte de sedimentos (YUAN et al., 2019).

2.3.2 Variáveis utilizadas na predição da concentração superficial de sedimentos

Dentre as variáveis preditoras utilizadas na modelagem da CSS, a dinâmica das classes de uso e cobertura da terra impactam em grande medida a produção de sedimentos em bacias hidrográficas. As diversas atividades antropogênicas, em especial as atividades agropecuárias, muitas vezes sem adoção de práticas de conservacionistas, resultam em aumento dos processos

erosivos, os quais potencializam a produção de sedimentos (RAJBANSHI; BHATTACHARYA, 2020).

A bacia hidrográfica do rio Doce é caracterizada pela exploração agropecuária de longa data, com uso antrópico bastante consolidado. Grande parte das pastagens encontram-se bastante degradadas, proporcionando pouca proteção ao solo (ECOPLAN-LUME, 2010). Nos Apêndices J e K é possível verificar a evolução das classes de uso e cobertura da terra nas regiões definidas como alto rio Doce e baixo rio Doce da bacia hidrográfica do rio Doce, para fins de modelagem da CSS, em cada ano entre 1985 e 2019. Na Figura 2.3 é apresentada a evolução das classes de uso e cobertura da terra para essas regiões nos anos 1985 e 2019.

Na Figura 2.3 é importante destacar a redução das áreas de pastagens em ambas as regiões analisadas, com 15% (1369,6 km²) e 17% (6600,5 km²), respectivamente. As pastagens ocupam a maior parte da bacia hidrográfica do rio Doce e essas reduções ocorreram principalmente devido ao aumento das áreas de agricultura e floresta plantada, que na região alto rio Doce tiveram acréscimo de 34% (418 km²) e 27%, respectivamente. Já na região baixo rio Doce, esses aumentos foram de 26% (1395,8 km²) nas áreas de agricultura e 25% (1551,0 km²) na composição das áreas de florestas plantadas.

A classe de uso da terra mosaico de agricultura e pastagem, que também apresentou aumento entre 1985 e 2019, representa áreas em que não foi possível a distinção entre agricultura e pastagem (SOUZA et al., 2020). Como as áreas de agricultura têm aumentando ao longo do tempo, é possível que o aumento observado nesta classe seja motivado principalmente pela agricultura.

As classes de usos da terra compostas por áreas não vegetadas, afloramento rochoso e corpos de água, apesar de apresentarem variações altas entre 1985 e 2019, são menos expressivas nas regiões analisadas. Essas variações podem estar relacionadas à melhor capacidade de detecção de alvos do satélite *Operational Land Imager* (OLI)/Landsat 8, com resolução radiométrica de 12 bits, em comparação com o satélite *Thematic Mapper* (TM)/ Landsat 5 (8 bits) (POURSANIDIS; CHRYSOULAKIS; MITRAKA, 2015).

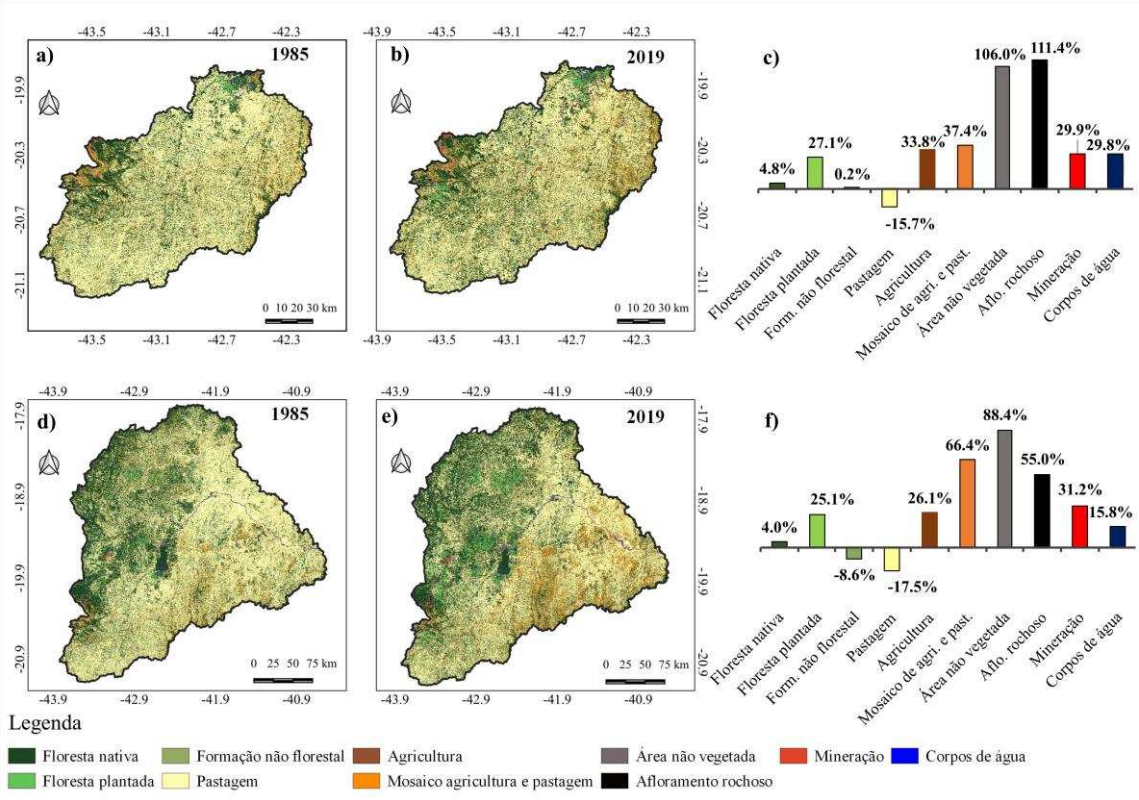
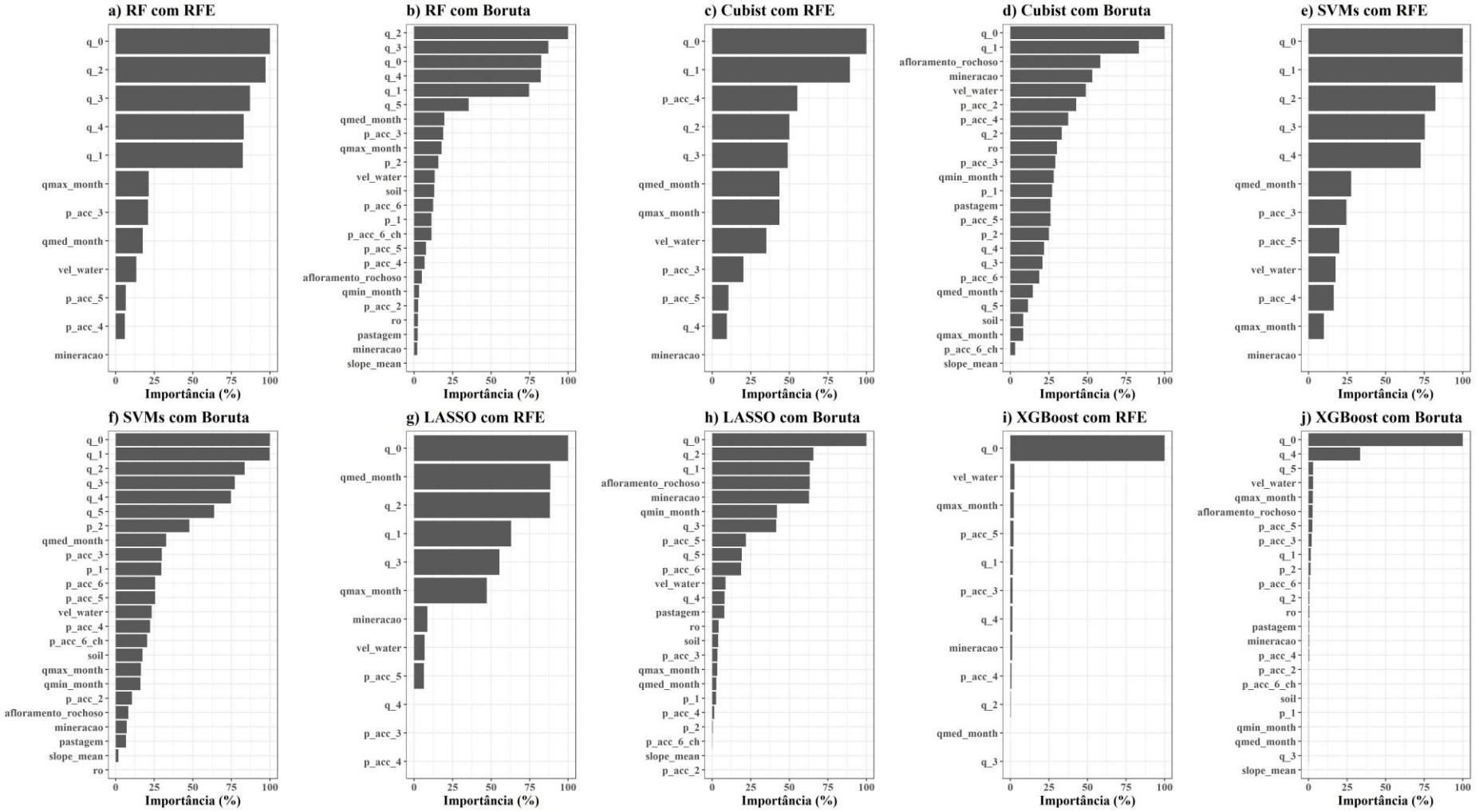


Figura 2.3. Mudanças das classes uso e cobertura da terra entre 1985 e 2019: (a) classes de usos da terra na região alto rio Doce em 1985 e (b) 2019; (c) variação das classes de usos da terra na região alto rio Doce entre 1985 e 2019; (d) classes de usos da terra na região baixo rio Doce em 1985 e (d) 2019; (f) variação das classes de usos da terra na região baixo rio Doce entre 1985 e 2019.

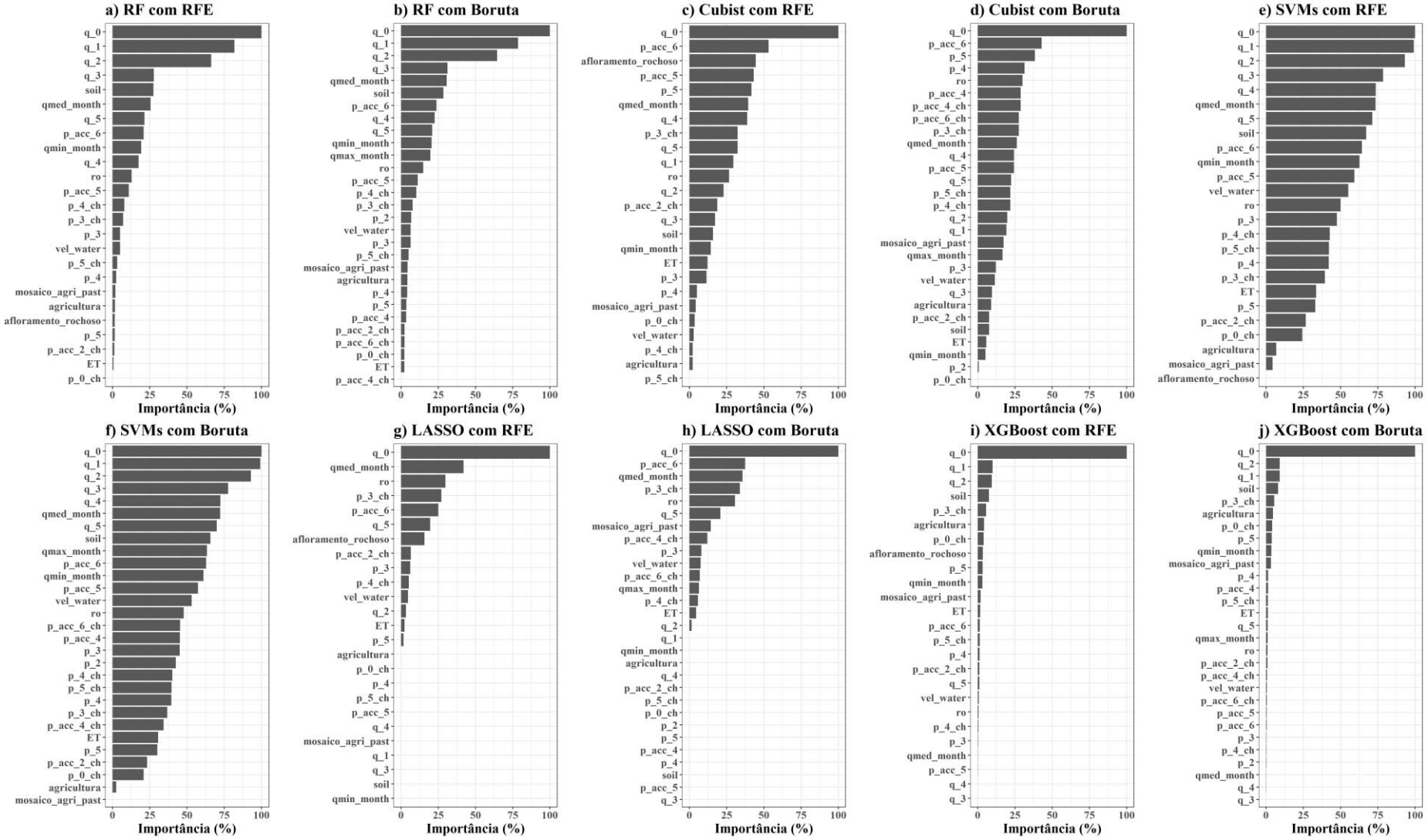
2.3.3. Seleção das variáveis utilizadas nos modelos de aprendizado de máquina

A partir da aplicação dos algoritmos de seleção de variáveis Boruta e *Recursive Feature Elimination* (RFE), o número de variáveis preditoras utilizadas nos modelos de aprendizado de máquina foi reduzido. Isso é fundamental para a otimização do processo, pois reduz o tempo de processamento e, além disso, permite a eliminação de variáveis que apresentam pouca relação com a variável resposta e que, em alguns casos, prejudicam a capacidade preditiva dos modelos (WEI et al., 2020).

Por meio da função *varImp* do pacote Caret, implementado no software R, obteve-se o ranqueamento em porcentagem da importância de cada variável em relação a sua capacidade minimizar os erros de predição (KUHN, 2008). Nas Figuras 2.4 e 2.5 apresenta-se as variáveis que foram selecionadas pelos algoritmos Boruta e RFE em cada modelo de aprendizado de máquina utilizado neste estudo.



Figuras 2.4. Importância das variáveis selecionadas em cada modelo de aprendizado de máquina para a região alto rio Doce da bacia hidrográfica do rio Doce



Figuras 2.5. Importância das variáveis selecionadas em cada modelo de aprendizado de máquina para a região baixo rio Doce da bacia hidrográfica do rio Doce.

Verifica-se que na região alto rio Doce o algoritmo RFE considerou 12 variáveis como sendo ótimas, enquanto por meio do Boruta 24 variáveis foram selecionadas. Na região do baixo rio Doce ambos os algoritmos selecionaram quantidades próximas de variáveis, sendo que o RFE indicou 25 e o Boruta 29, confirmando a importância da pré-seleção das variáveis, pois foi possível reduzir consideravelmente o número de variáveis iniciais que eram 62.

O grau de importância de cada variável foi bastante diferente em cada modelo analisado, entretanto, é possível observar que a variável vazão diária do dia da coleta de sedimentos (q_0) foi eleita como a mais importante em quase todos os modelos analisados, com exceção do modelo RF-Boruta (Figura 2.4 (b)) na região alto rio Doce, que classificou como mais importante a variável vazão diária defasada em 2 dias anteriores à coleta de sedimentos (q_2).

A maneira mais comum de monitorar a concentração superficial de sedimentos é por meio da obtenção da curva chave de sedimentos que, em geral, relaciona a vazão diária com a concentração de sedimentos (ZHENG, 2018; JAIYEOLA; ADEYEMO, 2019; KEOGH et al., 2019; BENISI GHADIM et al., 2020). Portanto, esperava-se que a vazão diária apresentasse boa relação com a CSS na área de estudos.

As vazões defasadas no tempo são, frequentemente, consideradas nos estudos sedimentométricos. A relação entre o aumento na CSS e a vazão fluvial apresenta atraso, ou seja, após o aumento da vazão ainda leva um certo tempo para que a CSS comece a aumentar, fenômeno conhecido como histerese (HAMSHAW et al., 2018; KEESSTRA et al., 2019; MALUTTA et al., 2020; CAO et al., 2021; HADDADCHI; HICKS, 2021). É possível observar nas Figuras 2.4 e 2.5 que em grande parte dos modelos as vazões defasadas do tempo se configuram entre as cinco variáveis mais importantes para a predição da CSS na área de estudo.

As precipitações médias diárias, principalmente as medidas pelos pluviômetros, também se destacaram nos modelos de predição da CSS. As precipitações médias diárias acumuladas em até 120h antes da coleta de sedimentos (p_acc_6) destacaram-se em grande parte dos modelos, em especial com o algoritmo Cubist na região baixo rio Doce, Figuras 2.5 (c) e (d), em que essa variável foi a segunda mais importante para o modelo. Os dados de precipitação são mais difíceis de apresentarem boa correlação com os dados de sedimentos, uma vez que dependem de diversos fatores, como por exemplo, a capacidade de infiltração de água no solo, tipo de cobertura vegetal, declividade da área, entre outras características (ZHAO et al., 2019).

As diversas variáveis de classes de uso e cobertura da terra, apesar de serem de grande relevância na produção de sedimentos em bacias hidrográficas, apresentaram importância limitada. Para a região alto rio Doce, utilizando o algoritmo de seleção de variáveis RFE, apenas a classe mineração foi considerada, já utilizando o Boruta as classes afloramento rochoso,

pastagem e mineração foram selecionadas. Na região baixo rio Doce, o RFE selecionou as classes de afloramento rochoso, agricultura e mosaico de agricultura e pastagem. Já o algoritmo Boruta selecionou as classes agricultura e mosaico de agricultura e pastagem. Como pode ser observado na Figura 2.3, as classes de agricultura e pastagem vem sofrendo alterações nas áreas analisadas, o que pode estar influenciando na produção de sedimentos nas áreas de estudo.

Variáveis como pedologia e declividade não foram selecionadas pelos algoritmos, as quais variam apenas em relação à área de drenagem de cada estação sedimentométrica. Portanto, variam muito pouco ao longo do tempo, já que os valores vão se repetindo à medida que o dado de determinada estação é selecionado.

Cada modelo de aprendizado de máquina fez o ranqueamento das variáveis de forma diferente. Isso depende muito da formulação matemática de cada algoritmo, sendo que em algumas situações variáveis pouco relacionadas fisicamente com a produção de sedimentos, como é o caso das áreas de afloramento rochoso, que são relativamente pouco expressivas na área de estudo, foram selecionadas. Isso porque as equações matemáticas derivadas da análise dificilmente podem ser relacionadas com o processo físico do fenômeno estudado, configurando-se numa das maiores críticas da utilização desses algoritmos, considerados assim como modelos caixa preta (CHEN; CHAU, 2016; ÖZGER; KABATAŞ, 2015). Por outro lado, por se tratar de modelagem complexa e não linear, os modelos de aprendizado de máquina são adequados à modelagem da CSS.

3.3.4 Avaliação dos modelos de aprendizado de máquina utilizados na predição da CSS

Os resultados das métricas de avaliação dos modelos para cada conjunto de variáveis selecionadas por meio do RFE e do Boruta podem ser analisadas na Figura 2.6, para a região alto rio Doce e Figura 2.7 para a região baixo rio Doce.

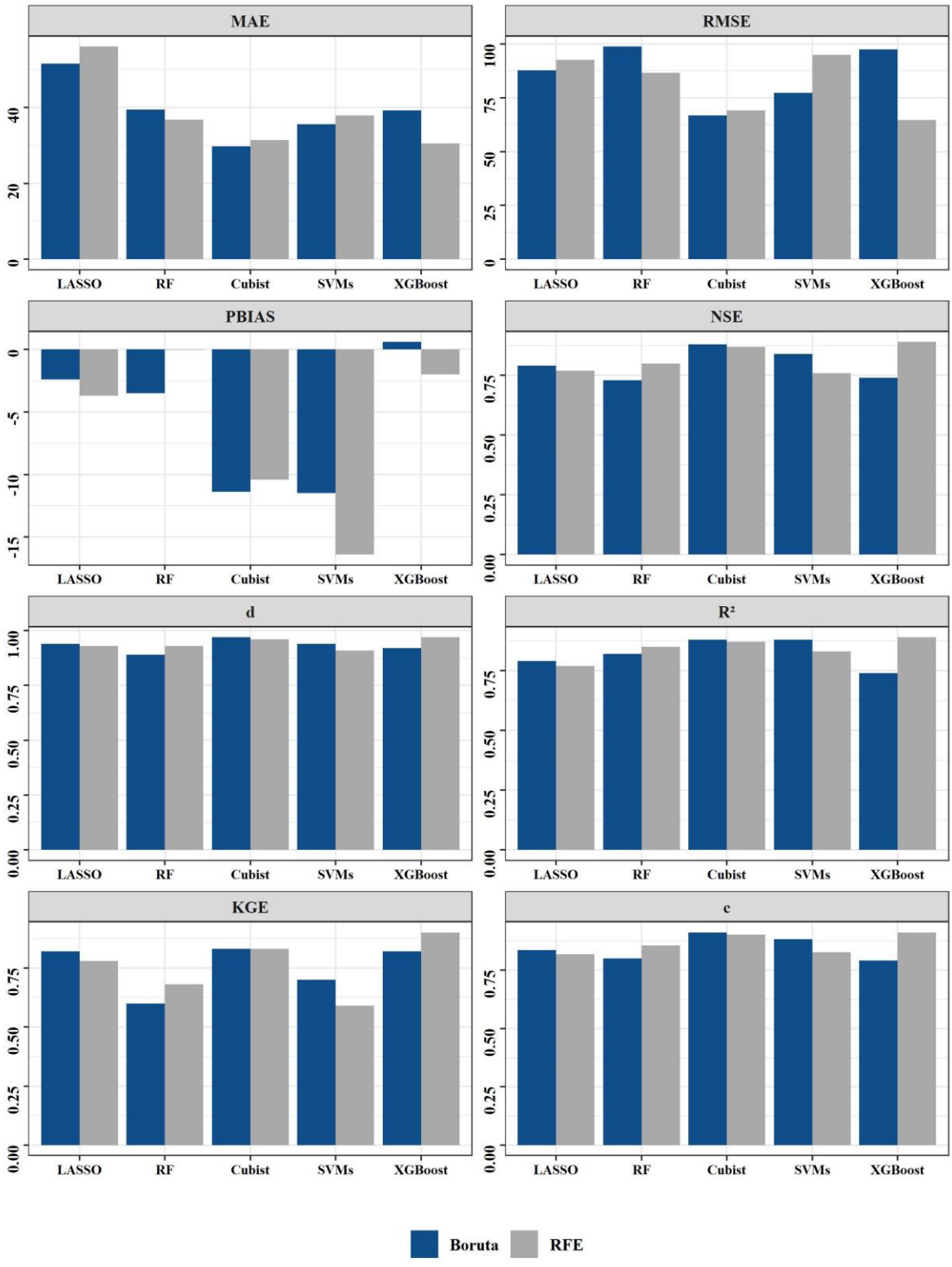


Figura 2.6. Métricas de avaliação dos modelos de aprendizado de máquina usados na modelagem da CSS para o conjunto de teste na região alto rio Doce da bacia hidrográfica do rio Doce.

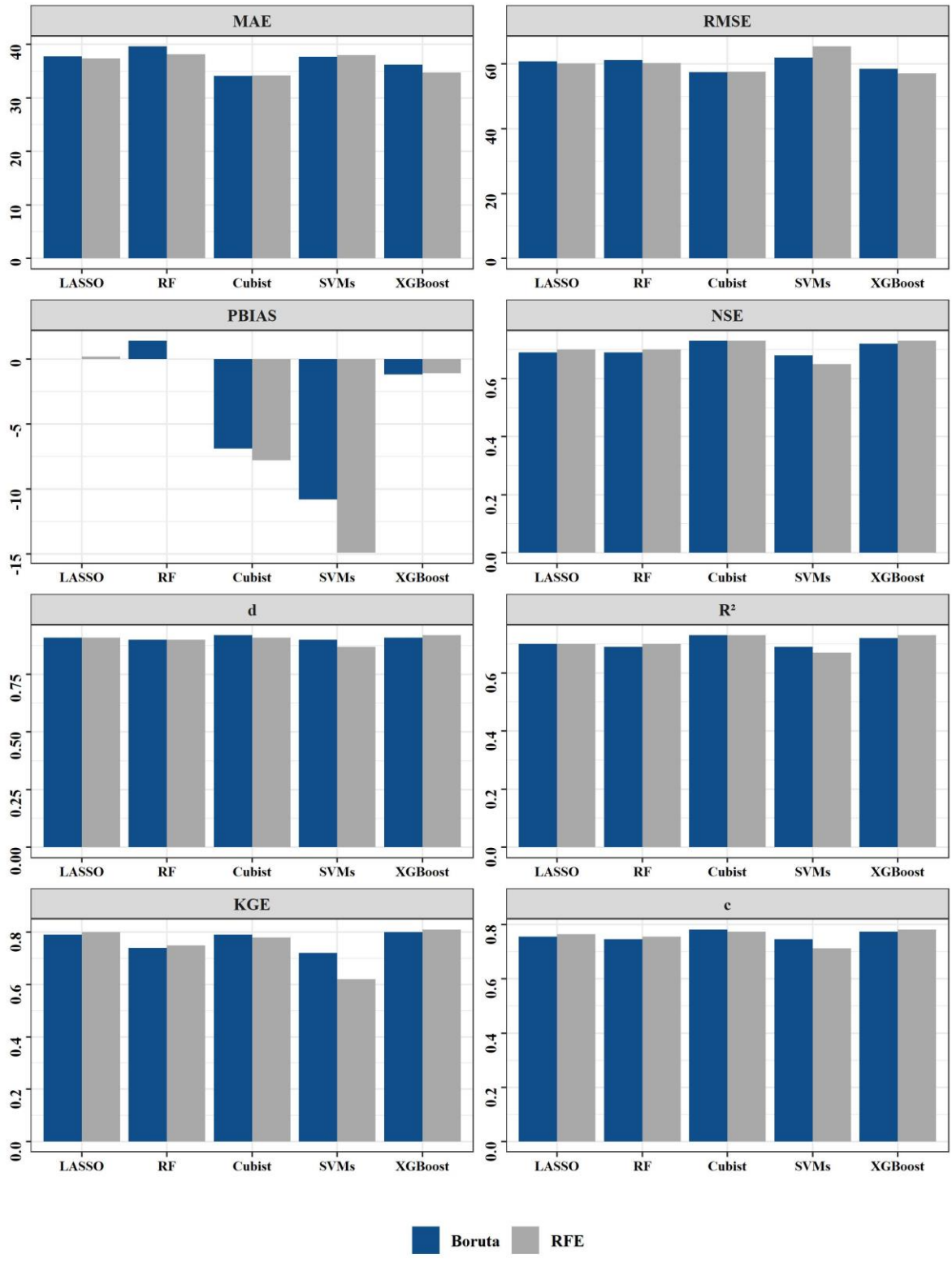


Figura 2.7. Métricas de avaliação dos modelos de aprendizado de máquina usados na modelagem da CSS para o conjunto de teste na região baixo rio Doce da bacia hidrográfica do rio Doce.

Como pode ser observado na Figura 2.6, referente à região definida alto rio Doce da bacia hidrográfica do rio Doce, os algoritmos de aprendizado de máquina Cubist-RFE e XGBoost-Boruta destacaram-se por apresentarem os menores valores de RMSE, com 66,73 mg/L e 64,57 mg/L, respectivamente. Esses dois modelos também se sobressaíram na região do baixo rio Doce (Figura 2.7), em que os valores de RMSE foram 57,4 mg/L para o modelo Cubist e 57 mg/L para o modelo XGboost-Boruta. Não ocorreram diferenças ao utilizar o conjunto de variáveis selecionadas pelo RFE ou Boruta para o modelo Cubist nesta região.

A maior parte dos modelos subestimaram os dados observados da CSS, com valores negativos de Pbias, tanto na região alto rio Doce como na região baixo rio Doce. O SVMs-RFE apresentou as maiores subestimativas, com valores de Pbias superiores a 15% na região alto rio Doce. O algoritmo Cubist-RFE também tendeu a subestimar os dados da CSS, com Pbias de até 11% na região baixo rio Doce. Valores de Pbias inferiores a 15% indicam que os modelos são satisfatórios, entretanto, o ideal é a obtenção de valores de Pbias inferiores a 5% (FERREIRA et al., 2021).

Dentre os índices que avaliam o desempenho dos modelos, o NSE e o KGE tem sido amplamente utilizados na avaliação dos modelos hidrológicos, sendo o KGE mais recomendo (RITTER; MUÑOZ-CARPENA, 2013; ALTHOFF; RODRIGUES, 2021). Neste sentido, na região alto rio Doce os valores de NSE mais elevados foram obtidos com os algoritmos Cubist-Boruta e XGBoost-RFE, com 0,88 e 0,89, respectivamente. Enquanto para o KGE o modelo XGBoost-RFE obteve o maior valor, com 0,9.

Na região baixo rio Doce os índices, apesar de serem satisfatórios, apontam que o ajuste dos modelos foi inferior em comparação com a região alto rio Doce. Em geral, todos os modelos utilizados apresentaram o NSE próximos, com valores entre 0,65 e 0,73. O KGE apresentou valores mais elevados em comparação com o NSE, em que o XGBoost-RFE obteve o maior valor (KGE = 0,81). Os algoritmos regressão LASSO e Cubist apresentaram valores muito similares, com KGE variando de 0,78 a 0,8.

Como pode ser analisado nas Figuras 2.6 e 2.7, os melhores resultados foram obtidos para a região alto rio Doce da bacia hidrográfica do rio Doce. Isso pode estar relacionado ao fato de a área ser bem menor em comparação com a região baixo rio Doce. Além disso, as estações sedimentométricas estão mais próximas, portanto, é esperado que os valores de sedimentos medidos apresentem maior homogeneidade.

Nas Figuras 2.8 e 2.9 é possível observar a dispersão dos dados preditos em relação aos dados observados para as regiões a alto rio Doce e baixo rio Doce, respectivamente,

considerando cada algoritmo de aprendizado de máquina utilizado na modelagem da CSS na bacia do rio Doce.

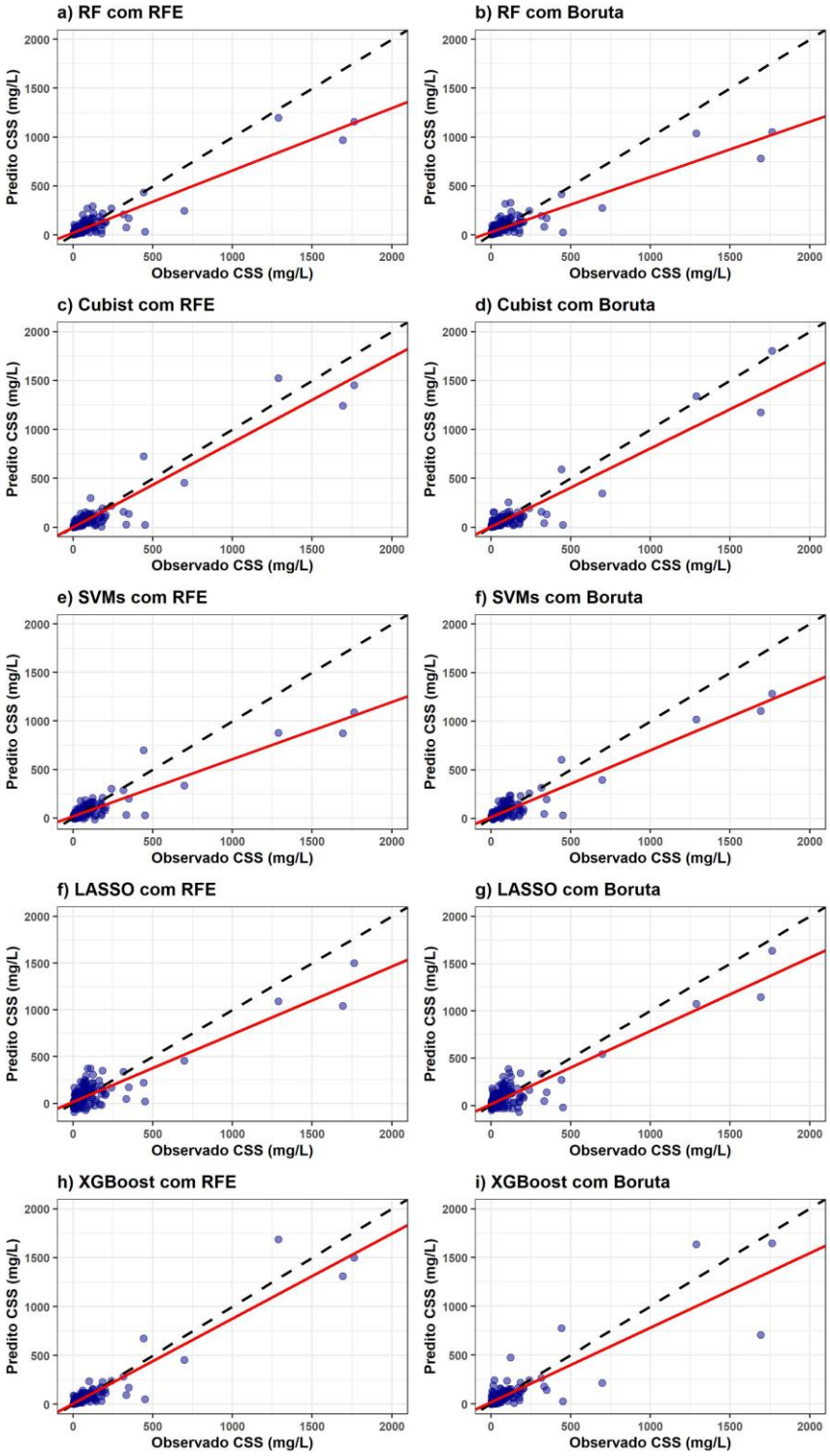


Figura 2.8. Relação dos valores preditos na modelagem da CSS e observados para a região alto rio Doce da bacia hidrográfica do rio Doce, em cada um dos algoritmos de aprendizado de máquina utilizados.

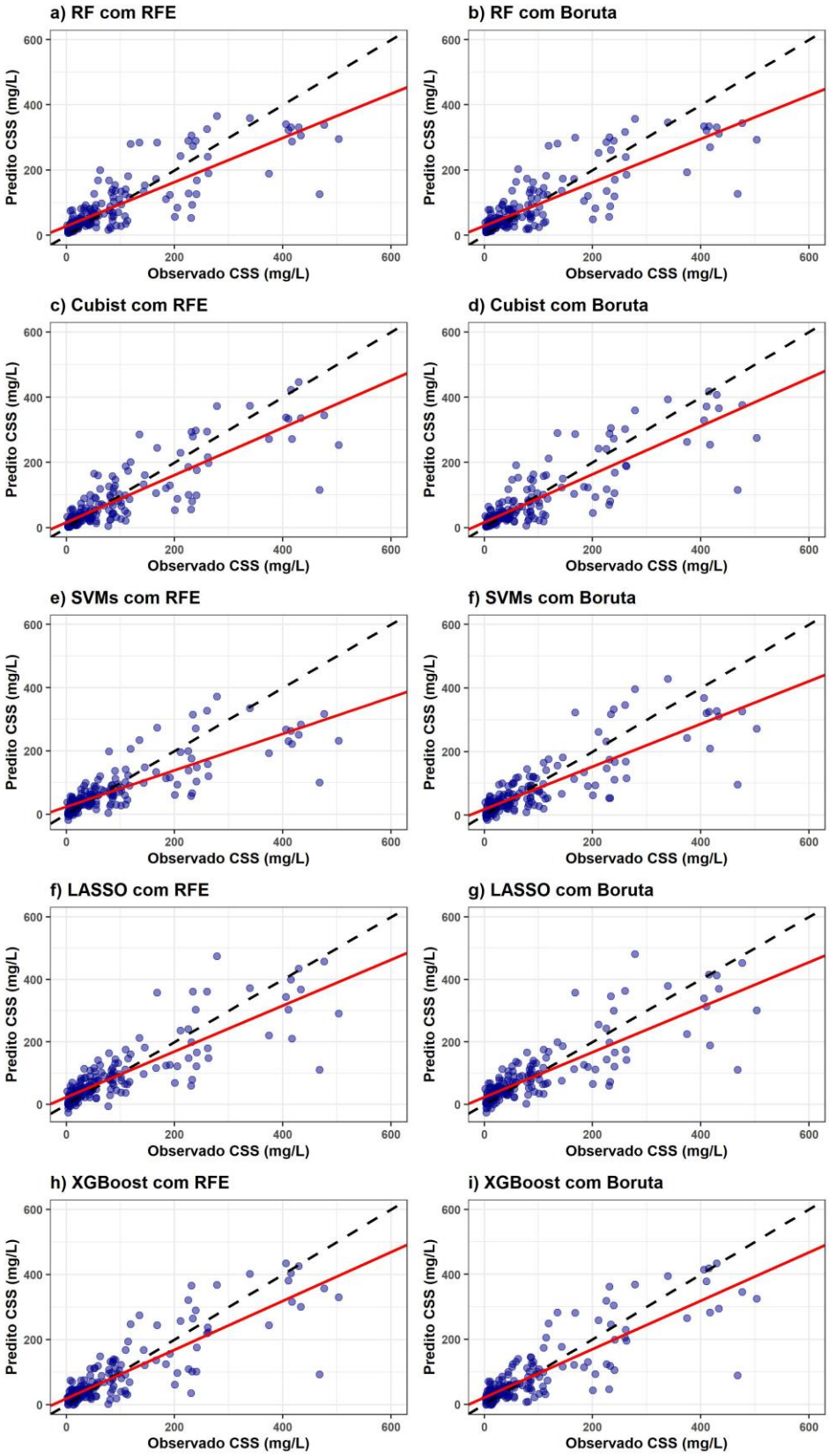


Figura 2.9. Relação dos valores preditos na modelagem da CSS e observados para a região baixo rio Doce da bacia hidrográfica do rio Doce, em cada um dos algoritmos de aprendizado de máquina utilizados.

Analisando a Figura 2.8, observa-se que os dados de CSS preditos pelos modelos Cubist-RFE e XGBoost-RFE foram os que mais se aproximaram da reta 1:1, o que indica menor dispersão entre os dados preditos e observados. Os dados preditos pelos modelos de regressão LASSO e SVMs apresentaram números negativos de CSS, o que não tem significado físico.

Os ajustes dos modelos para a região do baixo rio Doce foram inferiores àqueles obtidos na região alto rio Doce, resultando em valores mais dispersos em relação a reta 1:1 (Figura 2.9). Os algoritmos que mais se aproximaram da reta 1:1 foram o Cubist e o XGBoost, independente do algoritmo de seleção de variáveis utilizado. Os modelos SVMs, regressão LASSO e o XGBoost-RFE apresentaram valores preditos negativos para alguns casos, principalmente com valores baixos da CSS.

A utilização de algoritmos de aprendizado de máquina para o monitoramento da concentração superficial de sedimentos apresenta grande potencial, como observado nos resultados deste trabalho. Apesar de serem de difícil interpretação, diversos estudos tem demonstrado que esses modelos tem boa capacidade de trabalhar a variabilidade dos dados de sedimentos, que muitas vezes não apresenta relação direta com os dados de vazão fluvial (OLYAIE et al., 2015; RAMEZANI; NIKOO, 2015; CHEN; CHAU, 2016; KISI; ZOUNEMAT-KERMANI, 2016; BUYUKYILDIZ; KUMCU, 2017; UMAR; RHOADS; GREENBERG, 2018; PETERSON et al., 2018; AL-MUKHTAR; AL-YASEEN, 2019; SAMET et al., 2019; TAO; KESHTEGAR; YASEEN, 2019; YAWAR et al., 2019; HAMAAMIN et al., 2019; KISI; MUNDHER, 2019; SABERIOON et al., 2020).

De maneira geral, os modelos que apresentaram os melhores desempenhos foram o Cubist e o XGBoost. Entretanto, esse último demanda capacidade de processamento muito maior do que o Cubist. No estudo de Saberioon et al. (2020), bons resultados na predição da CSS também foram obtidos por meio do modelo Cubist, com valores de R^2 de 0,8 e RMSE de 19,55 mg/L, o que demonstrando o potencial desse algoritmo em estudos sedimentométricos.

É importante destacar a necessidade da utilização de diversos modelos, como foi o caso deste estudo, pois diferentes resultados são alcançados em cada algoritmo e a escolha da melhor opção irá depender do objetivo da pesquisa. Como exemplos, a redução dos erros de predição ou aquele modelo que proporciona erros aceitáveis, porém demandando menor capacidade de processamento.

A grande dificuldade da utilização dos modelos de aprendizado de máquina em hidrossedimentologia consiste na escassez de dados observados para o treinamento e teste dos modelos. No caso deste trabalho, foram agregadas diversas estações objetivando aumentar a base de dados. Entretanto, muitas bacias hidrográficas apresentam poucas estações

sedimentométricas e, em muitos casos, distantes uma das outras, o que aumenta muito a variabilidade dos dados. Isso pode ser observado no ajuste da região baixo rio Doce da bacia hidrográfica do rio Doce, que apresentou piores ajustes dos modelos em relação à região alto rio Doce.

2.4. CONCLUSÕES

Com base nos resultados obtidos neste estudo conclui-se que:

- A utilização de algoritmos de aprendizado de máquina para a predição da concentração superficial de sedimentos (CSS) na bacia hidrográfica do rio Doce apresentou bons resultados, com destaque para os modelos Cubist e XGBoost, que apresentaram o menor erro de predição e métricas de eficiência mais elevadas;
- A maioria dos modelos de aprendizado de máquina utilizados subestimaram os valores observados de CSS, inclusive os modelos LASSO e SVMs predizeram valores negativos;
- As variáveis mais importantes para os modelos de predição se configuraram nas vazões fluviais diárias na data da coleta e as vazões defasadas no tempo. A precipitação média diária acumulada observada nos pluviômetros também foi considerada importante na modelagem dos sedimentos;
- Os algoritmos de seleção de variáveis utilizados RFE e Boruta reduziram em grande parte o número de variáveis preditoras na modelagem da CSS, entretanto, não se verificou grandes diferenças na seleção das variáveis por ambos os algoritmos.

REFERÊNCIAS

ABATZOGLOU, J. T. et al. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. **Scientific Data 2018 5:1**, v. 5, n. 1, p. 1–12, 9 jan. 2018.

AFAN, H. A. et al. Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction. **Journal of Hydrology**, v. 541, p. 902–913, 1 out. 2016.

AIRES, U. R. V. et al. Changes in land use and land cover as a result of the failure of a mining tailings dam in Mariana, MG, Brazil. **Land Use Policy**, v. 70, 2018.

AL-MUKHTAR, M. Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. **Environmental Monitoring and Assessment**, v. 191, n. 11, p. 673, 25 nov. 2019.

AL-MUKHTAR, M.; AL-YASEEN, F. Modeling Water Quality Parameters Using Data-Driven Models, a Case Study Abu-Ziriq Marsh in South of Iraq. **Hydrology**, v. 6, n. 1, p. 24, 17 mar. 2019.

ALTHOFF, D.; RODRIGUES, L. N. Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. **Journal of Hydrology**, v. 600, p. 126674, 1 set. 2021.

ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 1 dez. 2013.

BENISI GHADIM, H. et al. Developing a Sediment Rating Curve ModelUsing the Curve Slope. **Polish Journal of Environmental Studies**, v. 29, n. 2, p. 1151–1159, 16 jan. 2020.

BHARTI, B. et al. Modelling of runoff and sediment yield using ANN , LS-SVR , REPTree and M5 models. **Hydrology Research**, p. 1489–1507, 2017.

BHATTACHARYA, B.; PRICE, R. K.; SOLOMATINE, D. P. Machine Learning Approach to Modeling Sediment Transport. **Journal of Hydraulic Engineering**, v. 133, n. 4, p. 776–793, 2007.

BUTLER, B. M.; O'ROURKE, S. M.; HILLIER, S. Using rule-based regression models to predict and interpret soil properties from X-ray powder diffraction data. **Geoderma**, v. 329, p. 43–53, nov. 2018.

BUYUKYILDIZ, M.; KUMCU, S. Y. An Estimation of the Suspended Sediment Load Using Adaptive Network Based Fuzzy Inference System , Support Vector Machine and Artificial Neural Network Models. p. 1343–1359, 2017.

CAMARGO, ângelo P. de; SENTELHAS, P. C. Avaliação do desempenho de diferentes métodos de estimativa da evapotranspiração potencial no Estado de São Paulo. **Revista Brasileira de Agrometeorologia**, v. 5, n. 1, p. 89–87, 1997.

CAO, L. et al. Factors controlling discharge-suspended sediment hysteresis in karst basins, southwest China: Implications for sediment management. **Journal of Hydrology**, v. 594, p. 125792, 1 mar. 2021.

CARVALHO, N. de O. et al. **Guia de práticas sedimentométricas**. Brasília: ANEEL, 2000.

CBH-DOCE. **A bacia hidrográfica do Rio Doce**. Disponível em: <http://www.cbhdoce.org.br/institucional/a-bacia>. Acesso em: 5 out. 2018.

CHEN, J. et al. Spatio-Temporal Patterns and Impacts of Sediment Variations in Downstream of the Three Gorges Dam on the Yangtze River , China. **Sustainability**, v. 10, n. 11, p. 1–17, 2018.

CHEN, T.; GUESTRIN, C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, **Anais**. 2016.

CHEN, X. Y.; CHAU, K. W. A Hybrid Double Feedforward Neural Network for Suspended Sediment Load Estimation. **Water Resources Management**, v. 30, n. 7, p. 2179–2194, 1 maio 2016.

CHIRPS. **CHIRPS: Rainfall Estimates from Rain Gauge and Satellite Observations | Climate Hazards Center - UC Santa Barbara**. Disponível em: <https://www.chc.ucsb.edu/data/chirps>. Acesso em: 29 jan. 2022.

ECOPLAN-LUME. **Plano Integrado de Recursos Hídricos da Bacia Hidrográfica do Rio Doce - PIRH Bacia do Rio Doce**. 1. ed. Belo Horizonte: CONSÓRCIO ECOPLAN-LUME, 2010.

EFTHIMIOU, N.; LYKOUDI, E.; KARAVITIS, C. Comparative analysis of sediment yield estimations using different empirical soil erosion models. **Hydrological Sciences Journal**, v. 00, n. 00, p. 1–21, 2017.

ELESBON, A. A. A. et al. Multivariate statistical analysis to support the minimum streamflow regionalization. **Engenharia Agrícola**, v. 35, n. 5, p. 838–851, out. 2015.

EMBRAPA. **Sistema Brasileiro de Classificação de Solos**. 5. ed. Brasília: Embrapa, 2018.

FARR, T. G. et al. The shuttle radar topography mission. **Reviews of Geophysics**, v. 45, n. 2, jun. 2007.

FERREIRA, R. G. et al. Machine learning models for streamflow regionalization in a tropical watershed. **Journal of Environmental Management**, v. 280, p. 111713, 15 fev. 2021.

FROMANT, G. et al. Suspended sediment concentration field quantified from a calibrated MultiBeam EchoSounder. **Applied Acoustics**, v. 180, p. 108107, 1 set. 2021.

FUNK, C. et al. The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes. **Nature: Scientific Data**, v. 2, p. 1–21, 2015.

GEE. **Introduction | Google Earth Engine API | Google Developers**. Disponível em: <https://developers.google.com/earth-engine/>. Acesso em: 13 dez. 2017.

GRANITTO, P. M. et al. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. **Chemometrics and Intelligent Laboratory Systems**, v. 83, n. 2, p. 83–90, 15 set. 2006.

GUPTA, H. V. et al. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. **Journal of Hydrology**, v. 377, n. 1–2, p. 80–91, 20 out. 2009.

HADDADCHI, A.; HICKS, M. Interpreting event-based suspended sediment concentration and flow hysteresis patterns. **Journal of Soils and Sediments**, v. 21, p. 592–612, 2021.

HAIR, J. F. et al. Análise de Regressão Múltipla. In: HAIR, J. F. et al. (Ed.). **Análise Multivariada de Dados**. 6. ed. Porto alegre: Bookman: Bookman, 2009. p. 149–220.

HALLAK, R.; PEREIRA FILHO, A. J. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS. **Revista Brasileira de Meteorologia**, v. 26, n. 4, p. 591–608, 2011.

HAMAAMIN, Y. A. et al. Evaluation of neuro-fuzzy and Bayesian techniques in estimating suspended sediment loads. **Sustainable Water Resources Management**, v. 5, n. 2, p. 639–654, 2019.

HAMSHAW, S. D. et al. A New Machine-Learning Approach for Classifying Hysteresis in Suspended-Sediment Discharge Relationships Using High-Frequency Monitoring Data. **Water Resources Research**, v. 54, n. 6, p. 4040–4058, 1 jun. 2018.

HEARST, M. A. et al. Support vector machines. In: IEEE Intelligent Systems and their Applications, 4, **Anais**. jul. 1998.

HIDROWEB. **Séries Históricas de Estações**. Disponível em: <https://www.snirh.gov.br/hidroweb/serieshistoricas>. Acesso em: 11 jan. 2022.

HIMANSHU, S. K.; PANDEY, A.; YADAV, B. Assessing the applicability of TMPA-3B42V7 precipitation dataset in wavelet-support vector machine approach for suspended sediment load prediction. **Journal of Hydrology**, v. 550, p. 103–117, 2017.

IBGE. **Downloads: Geociências**. Disponível em: <https://www.ibge.gov.br/geociencias/downloads-geociencias.html>. Acesso em: 7 jan. 2022.

JAIYEOLA, A. T.; ADEYEMO, J. Performance comparison between genetic programming and sediment rating curve for suspended sediment prediction. **https://doi.org/10.1080/20421338.2019.1587908**, v. 11, n. 7, p. 843–859, 10 nov. 2019.

JAMES, G. et al. **An Introduction to Statistical Learning**. 2. ed. New York: Springer, 2021.

KAVEH, K.; DUC BUI, M.; RUTSCHMANN, P. A comparative study of three different learning algorithms applied to ANFIS for predicting daily suspended sediment concentration. **International Journal of Sediment Research**, v. 32, n. 3, p. 340–350, 2017.

KEESSTRA, S. D. et al. Coupling hysteresis analysis with sediment and hydrological connectivity in three agricultural catchments in Navarre, Spain. **Journal of Soils and Sediments**, v. 19, n. 3, p. 1598–1612, 11 mar. 2019.

KEOGH, M. E. et al. Hydrodynamic controls on sediment retention in an emerging diversion-fed delta. **Geomorphology**, v. 332, p. 100–111, 1 maio 2019.

KISI, O.; MUNDHER, Z. The potential of hybrid evolutionary fuzzy intelligence model for suspended sediment concentration prediction. **Catena**, v. 174, p. 11–23, 2019.

KISI, O.; ZOUNEMAT-KERMANI, M. Suspended Sediment Modeling Using Neuro-Fuzzy Embedded Fuzzy c-Means Clustering Technique. **Water Resources Management**, p. 3979–3994, 2016.

KUHN, M. Building Predictive Models in R Using the caret Package. **Journal of Statistical Software**, v. 28, n. 5, p. 1–26, 10 nov. 2008.

KUHN, M. **Caret: Classification and Regression Training. R package version 6.0-89**. 2021. https://CRAN.R-project.org/package=caret

KURSA, M. B.; JANKOWSKI, A.; RUDNICKI, W. R. Boruta – A System for Feature Selection. **Fundamenta Informaticae**, v. 101, n. 4, p. 271–285, 1 jan. 2010.

LAFDANI, E. K.; NIA, A. M.; AHMADI, A. Daily suspended sediment load prediction using artificial neural networks and support vector machines. **Journal of Hydrology**, v. 478, p. 50–62, 2013.

LI, P. et al. Soil erosion rates assessed by RUSLE and PESERA for a Chinese Loess Plateau catchment under land-cover changes. **Earth Surface Processes and Landforms**, v. 45, n. 3, p. 707–722, 15 mar. 2020.

LI, Z. et al. Impacts of land use change and climate variability on hydrology in an agricultural catchment on the Loess Plateau of China. **Journal of Hydrology**, v. 377, n. 1–2, p. 35–42, 20 out. 2009.

MAGESH, N. S.; CHANDRASEKAR, N. Assessment of soil erosion and sediment yield in the Tamiraparani sub-basin , South India , using an automated RUSLE-SY model. **Environmental Earth Sciences**, v. 75, n. 16, p. 1–17, 2016.

MALIK, A.; KUMAR, A.; PIRI, J. Daily suspended sediment concentration simulation using hydrological data of Pranhita River Basin, India. **Computers and Electronics in Agriculture**, v. 138, p. 20–28, 2017.

MALUTTA, S. et al. Hysteresis analysis to quantify and qualify the sediment dynamics: state of the art. **Water Science and Technology**, v. 81, n. 12, p. 2471–2487, 15 jun. 2020

ZAMBRANO-BIGIARINI, M. **HydroGOF: Goodness-of-fit functions fo comparison of simulated and observed hydrological time seriesR package version 0.4-0**. 2020. URL https://github.com/hzambran/hydroGOF. DOI:10.5281/zenodo.839854.

MAPBIOMAS. **Mapas e dados**. Disponível em: <https://mapbiomas.org/colecoes-mapbiomas-1?cama_set_language=pt-BR>. Acesso em: 10 jan. 2022.

MELLO, C. R.; SILVA, A. M. **Hidrologia: princípios e aplicações em sistemas agrícolas**. 1. ed. Lavras: Editora UFLA, Lavras, 455p., 2013., 2013.

KURSA, M. B., RUDNICKI, W. R. Feature Selection with the Boruta Package. **Journal of Statistical Software**, 36(11), 1-13, 2010.

MUSTAFA, M. R. Modeling daily suspended sediments of a hyper-concentrated river in Malaysia. **ARPN Journal of Engineering and Applied Sciences**, v. 11, n. 4, p. 2141–2145, 2016.

NAGUETTINI, M.; PINTO, E. J. A. **Hidrologia Estatística**. 1. ed. Belo Horizonte: CPRM, 2007.

NI, L. et al. Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. **Journal of Hydrology**, v. 586, p. 124901, 1 jul. 2020.

NOROUZI, R.; DANESHFARAZ, R.; GHADERI, A. Investigation of discharge coefficient of trapezoidal labyrinth weirs using artificial neural networks and support vector machines. **Applied Water Science**, v. 9, n. 7, p. 1–10, 6 out. 2019.

NOURANI, V. et al. Applications of hybrid wavelet – Artificial Intelligence models in hydrology : A review. **Journal of Hydrology**, v. 514, p. 358–377, 2014.

NOURANI, V.; ALIZADEH, F.; ROUSHANGAR, K. Evaluation of a Two-Stage SVM and Spatial Statistics Methods for Modeling Monthly River Suspended Sediment Load. **Water Resour Manage**, n. 30, p. 393–407, 2016.

OLYAIE, E. et al. A comparison of various artificial intelligence approaches performance for estimating suspended sediment load of river systems: a case study in United States. **Environmental Monitoring and Assessment**, v. 187, n. 4, 2015.

ÖZGER, M.; KABATAŞ, M. B. Sediment load prediction by combined fuzzy logic-wavelet method. **Journal of Hydroinformatics**, v. 17, n. 6, p. 930–942, 11 nov. 2015.

PEDELTY, J. et al. Generating a long-term land data record from the AVHRR and MODIS instruments. **International Geoscience and Remote Sensing Symposium (IGARSS)**, p. 1021–1024, 2007.

PEREIRA, H. R. et al. On the performance of three indices of agreement: an easy-to-use r-code for calculating the Willmott indices. **Bragantia**, v. 77, n. 2, p. 394–403, 22 mar. 2018.

PETERSON, K. T. et al. Suspended Sediment Concentration Estimation from Landsat Imagery along the Lower Missouri and Middle Mississippi Rivers Using an Extreme Learning Machine. **Remote Sensing**, v. 10, n. 10, p. 1–17, 2018.

PIÑEIRO, G. et al. How to evaluate models: Observed vs. predicted or predicted vs. observed? **Ecological Modelling**, v. 216, n. 3–4, p. 316–322, 10 set. 2008.

POURSANIDIS, D.; CHRYSOULAKIS, N.; MITRAKA, Z. Landsat 8 vs. Landsat 5: A comparison based on urban and peri-urban land cover mapping. **International Journal of Applied Earth Observation and Geoinformation**, v. 35, n. PB, p. 259–269, 1 mar. 2015.

QUINLAN, J. R. Learning with continuous classes. In: Proceedings of the 5th Australian Joint Conference On Artificial Intelligence, Australian. **Anais**. Australian: Utgoff, 1992.

QGIS DEVELOPMENT TEAM. **QGIS Geographic Information System**. Open Source Geospatial Foundation Project. http://qgis.osgeo.org, 2021.

RAHGOSHAY, M. et al. Simulation of daily suspended sediment load using an improved model of support vector machine and genetic algorithms and particle swarm. **Arabian Journal of Geosciences**, v. 12, n. 277, p. 1–14, 2019.

RAJBANSHI, J.; BHATTACHARYA, S. Assessment of soil erosion, sediment yield and basin specific controlling factors using RUSLE-SDR and PLSR approach in Konar river basin, India. **Journal of Hydrology**, v. 587, p. 124935, 1 ago. 2020.

RAMEZANI, F.; NIKOO, M. Artificial neural network weights optimization based on social-based algorithm to realize sediment over the river. p. 375–387, 2015.

RASHIDI, S.; VAFAKHAH, M. Evaluating the support vector machine for suspended sediment load forecasting based on gamma test. **Arabian Journal of Geosciences**, v. 9, n. 583, p. 1–15, 2016. Disponível em: <http://dx.doi.org/10.1007/s12517-016-2601-9>.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing,Vienna, Austria. 2021. URL https://www.R-project.org/.

RESTREPO, J. D.; ESCOBAR, H. A. Sediment load trends in the Magdalena River basin (1980–2010): Anthropogenic and climate-induced causes. **Geomorphology**, v. 302, p. 76–91, 2018.

RICCI, G. F.; GIROLAMO, A. M. De. Identifying sediment source areas in a Mediterranean watershed using the SWAT model. **Land Degradation & Development**, p. 1233–1248, 2018.

RITTER, A.; MUÑOZ-CARPENA, R. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. **Journal of Hydrology**, v. 480, p. 33–45, 14 fev. 2013.

SABERIOON, M. et al. Chlorophyll-a and total suspended solids retrieval and mapping using Sentinel-2A and machine learning for inland waters. **Ecological Indicators**, v. 113, p. 106236, 1 jun. 2020.

SAMET, K. et al. Comparison Between Soft Computing Methods for Prediction of Sediment Load in Rivers : Maku Dam Case Study. **Iranian Journal of Science and Technology, Transactions of Civil Engineering**, v. 43, n. 1, p. 93–103, 2019.

SHIAU, J.; CHEN, T. Quantile Regression-Based Probabilistic Estimation Scheme for Daily and Annual Suspended Sediment Loads. **Water Resour Manage**, p. 2805–2818, 2015.

SOUZA, C. M. et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing 2020, Vol. 12, Page 2735**, v. 12, n. 17, p. 2735, 25 ago. 2020.

SRIVASTAVA, A. et al. Science of the Total Environment Modeling forest management effects on water and sediment yield from nested , paired watersheds in the interior Pacific Northwest , USA using WEPP. **Science of the Total Environment**, v. 701, p. 1–14, 2019.

TANIGUCHI, H.; SATO, H.; SHIRAKAWA, T. A machine learning model with human cognitive biases capable of learning from small and biased datasets. **Scientific Reports**, v. 8, n. 1, p. 7397, 9 dez. 2018.

TAO, H.; KESHTEGAR, B.; YASEEN, Z. M. The Feasibility of Integrative Radial Basis M5Tree Predictive Model for River Suspended Sediment Load Simulation. **Water Resources Management**, n. 33, p. 4471–4490, 2019.

TAVAKOLI TARGHI, A.; ABBASZADEH, S.; ARABASADI, Z. A hybrid method for forecasting river-suspended sediments in Iran. **International Journal of River Basin Management**, v. 15, n. 4, p. 453–460, 2017.

UMAR, M.; RHOADS, B. L.; GREENBERG, J. A. Use of multispectral satellite remote sensing to assess mixing of suspended sediment downstream of large river confluences. **Journal of Hydrology**, v. 556, p. 325–338, 2018.

WEI, G. et al. A novel hybrid feature selection method based on dynamic feature importance. **Applied Soft Computing**, v. 93, p. 106337, 1 ago. 2020.

YAWAR, M. et al. Arti fi cial neural network simulation for prediction of suspended sediment concentration in the River Ramganga , Ganges Basin , India. **International Journal of Sediment Research**, v. 34, n. 2, p. 95–107, 2019.

YESUF, H. M. et al. Catena Modeling of sediment yield in Maybar gauged watershed using SWAT , northeast Ethiopia. **Catena**, v. 127, p. 191–205, 2015.

YUAN, X. P. et al. A New Efficient Method to Solve the Stream Power Law Model Taking Into Account Sediment Deposition. **Journal of Geophysical Research: Earth Surface**, v. 124, n. 6, p. 1346–1365, 1 jun. 2019.

ZHAO, Y. et al. Analysis of changes in characteristics of flood and sediment yield in typical basins of the Yellow River under extreme rainfall events. **CATENA**, v. 177, p. 31–40, 1 jun. 2019.

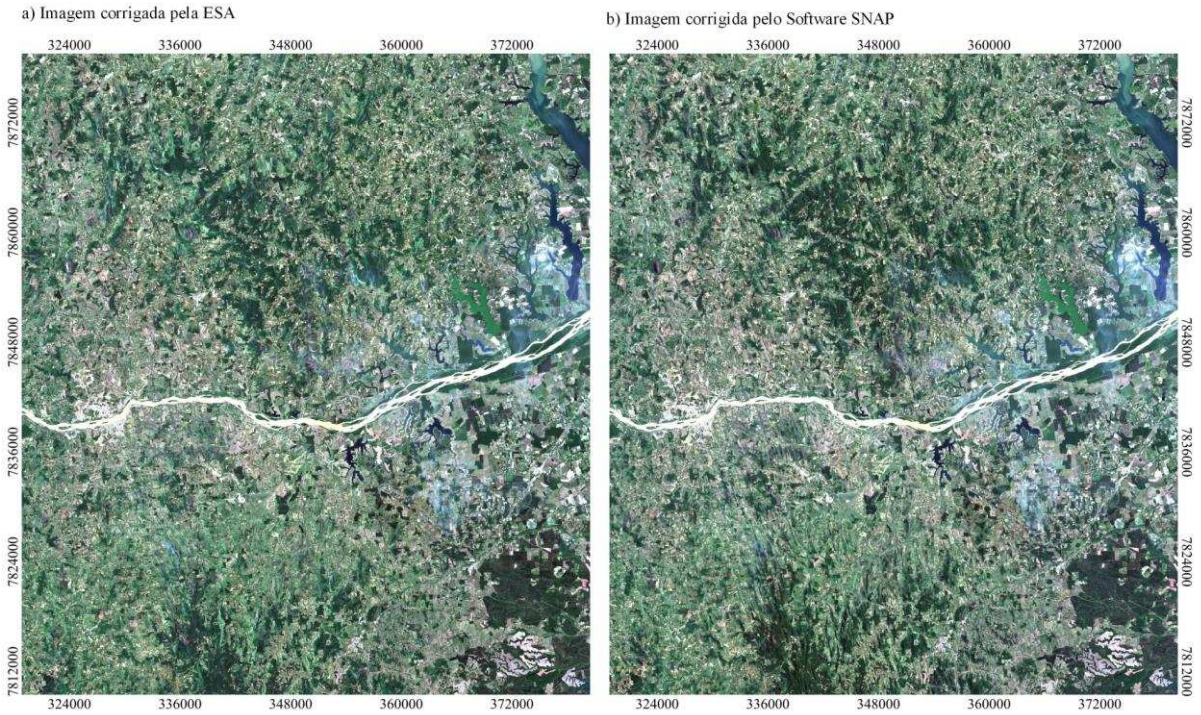
ZHENG, M. A spatially invariant sediment rating curve and its temporal change following watershed management in the Chinese Loess Plateau. **Science of The Total Environment**, v. 630, p. 1453–1463, 15 jul. 2018.

CONCLUSÕES GERAIS

- É possível realizar o monitoramento da concentração superficial de sedimentos (CSS) utilizando sensoriamento remoto orbital na calha do rio principal da bacia hidrográfica do rio Doce, por meio da relação linear entre a refletância medida pelo sensor orbital e os dados observados da CSS;
- A banda do infravermelho próximo apresentou forte relação linear com a CSS, tanto utilizando o satélite MSI/Sentinel 2 quanto o OLI/Landsat 8;
- As bandas do visível e infravermelho próximo (VNIRs) com 20 m de resolução espacial do satélite MSI/Sentinel 2, apresentaram boa relação linear no monitoramento da CSS, evidenciando o potencial desse satélite para o monitoramento dos recursos hídricos;
- Dentre os modelos de regressão linear que utilizam múltiplas variáveis, tanto a regressão linear múltipla quando a regressão LASSO e a regressão *Elastic Net* apresentaram bom desempenho para a predição dos sedimentos em suspensão, principalmente utilizando o Satélite MSI/Sentinel 2. Entretanto, estas últimas facilitam na definição do conjunto ótimo de variáveis;
- Os mapas de fluxos de sedimentos indicam redução da CSS na calha do rio Doce em anos mais recentes, o que pode ser indicativo de que parte do material oriundo do rompimento da barragem de rejeitos de Fundão pode ter sido carreado pelos processos de ressuspensão e transporte de sedimentos.
- A utilização de algoritmos de aprendizado de máquina para a predição da concentração superficial de sedimentos (CSS) na bacia hidrográfica do rio Doce apresentou bons resultados, com destaque para os modelos Cubist e XGBoost, que apresentaram o menor erro de predição e métricas de eficiência mais elevadas;
- A maior parte dos modelos de aprendizado de máquina subestimaram os valores observados de CSS, sendo que na regressão LASSO e no SVMs obteve-se, inclusive, valores negativos;
- As variáveis mais importantes para os modelos de predição se configuraram nas vazões fluviais diárias na data da coleta e as vazões defasadas no tempo. A precipitação média diária acumulada observada nos pluviômetros também foi considerada importante na modelagem dos sedimentos;
- Os algoritmos de seleção de variáveis utilizados RFE e Boruta reduziram em grande parte o número de variáveis preditoras na modelagem da CSS, entretanto, não se verificou grandes diferenças na seleção das variáveis por ambos algoritmos.

APÊNDICES

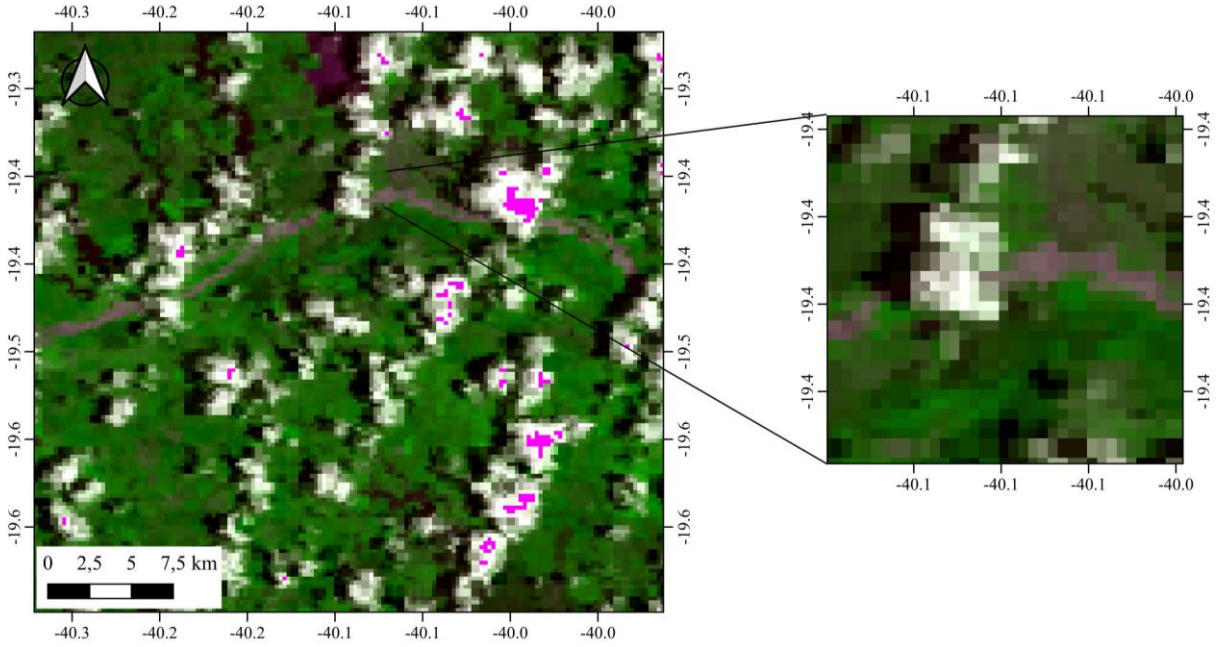
APÊNDICE A. Métricas de avaliação entre a imagem corrigida pela *European Space Agency* (ESA) e pelo software SNAP versão 8.0



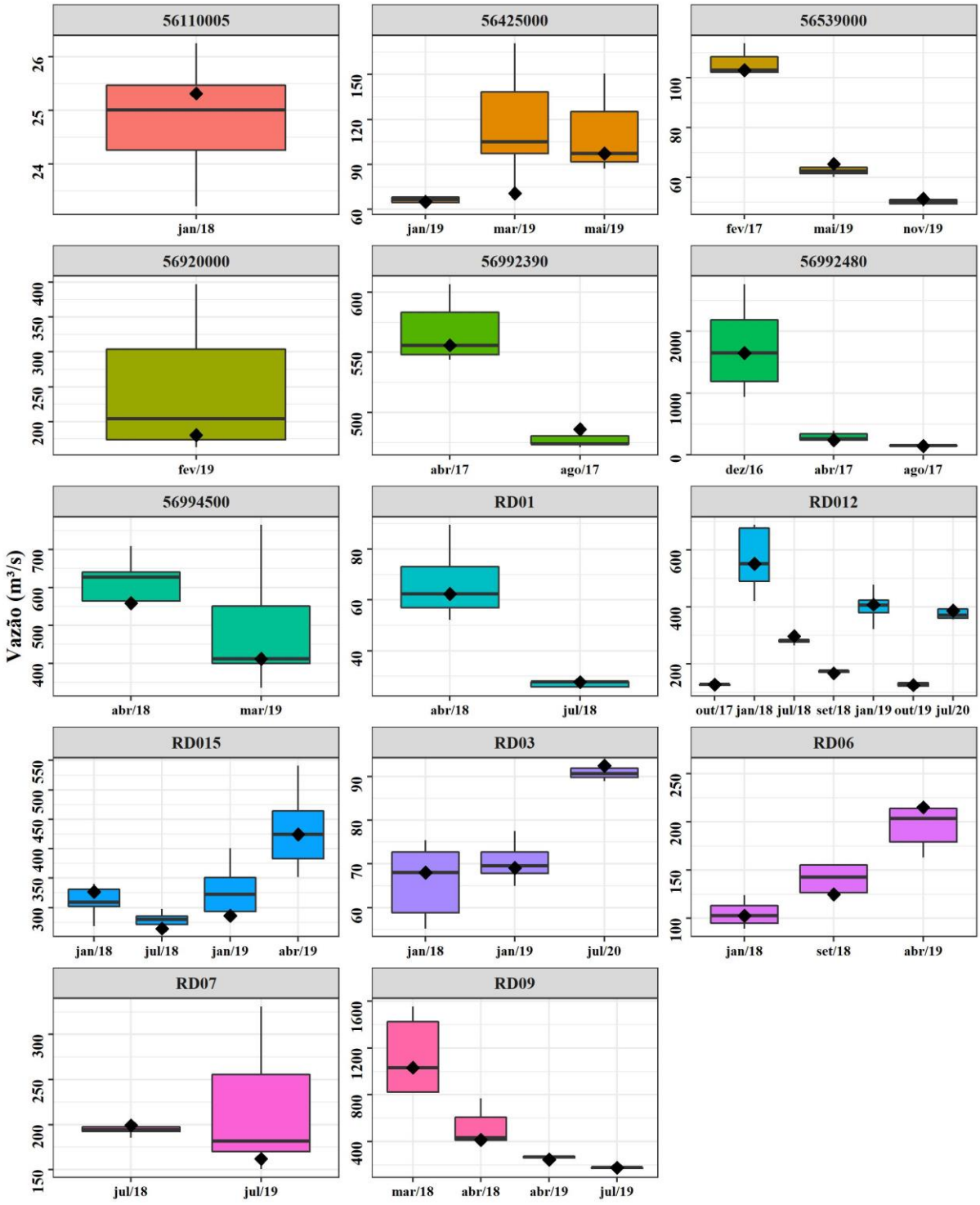
Bandas	MAE	RMSE	R^2	r	d
B02	0,003	0,004	0,95	0,98	0,99
B03	0,01	0,01	0,92	0,96	0,98
B04	0,004	0,01	0,98	0,99	0,99
B05	0,01	0,01	0,91	0,95	0,98
B06	0,02	0,03	0,75	0,87	0,93
B07	0,02	0,04	0,77	0,88	0,93
B08	0,02	0,04	0,78	0,88	0,94
B11	0,02	0,02	0,89	0,94	0,97
B12	0,01	0,01	0,95	0,97	0,99
B08A	0,03	0,04	0,77	0,88	0,93

Data da imagem: 25/02/2019; número de amostras: 2000

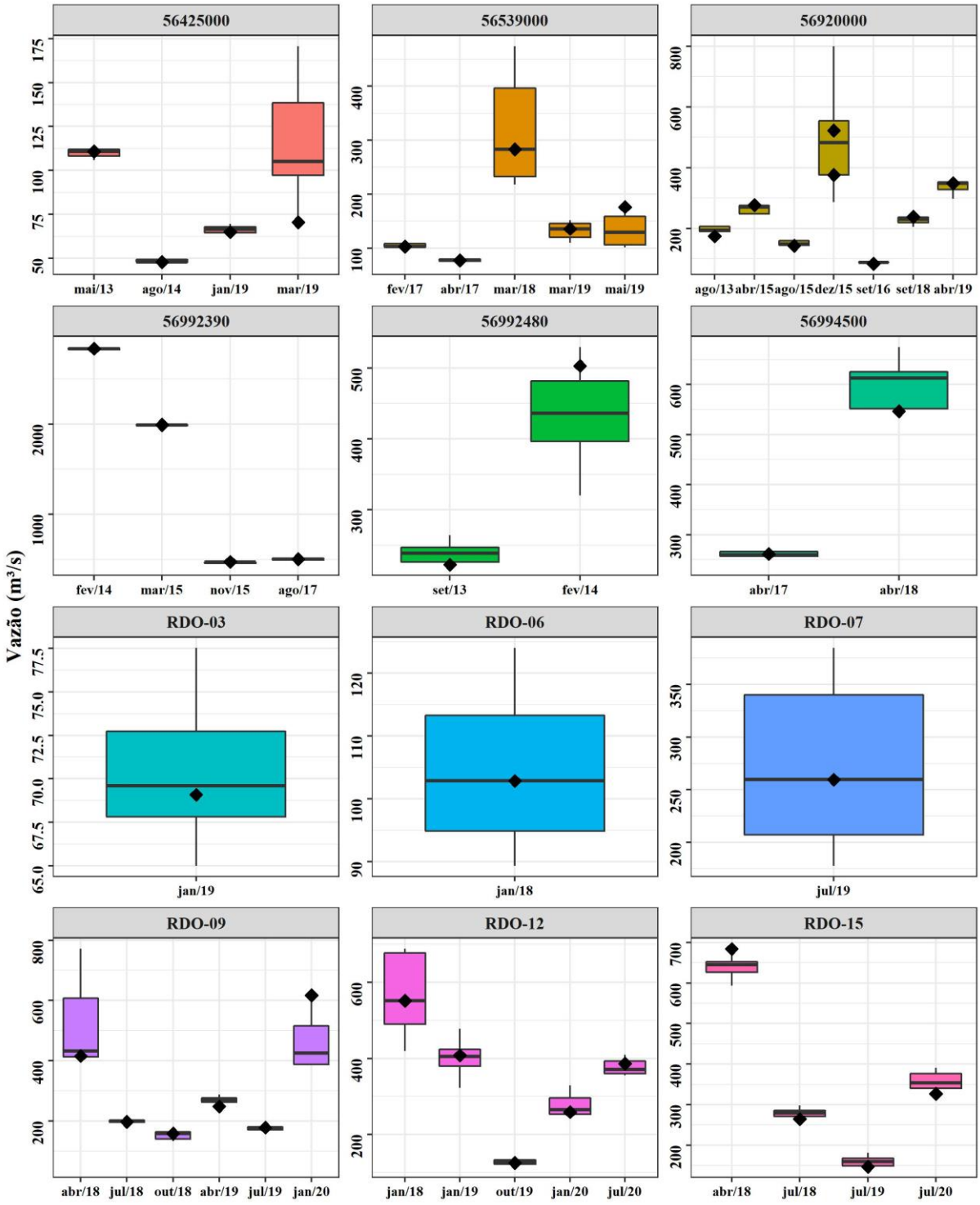
APÊNDICE B. Imagem do satélite MODIS/AQUA para o dia 15/12/2015 na região de Linhares, ES



APÊNDICE C. Variação da vazão nas estações sedimentométricas no período de três dias antes e após a data da coleta de sedimentos em relação aos dias em que foram possíveis a obtenção das imagens do satélite MSI/Sentinel 2. O ponto em formato de losango representa a vazão da data da coleta.



APÊNDICE D. Variação da vazão nas estações sedimentométricas no período de três dias antes e após a data da coleta de sedimentos em relação aos dias em que foram possíveis a obtenção das imagens do satélite OLI/Landsat 8. O ponto em formato de losango representa a vazão da data da coleta.



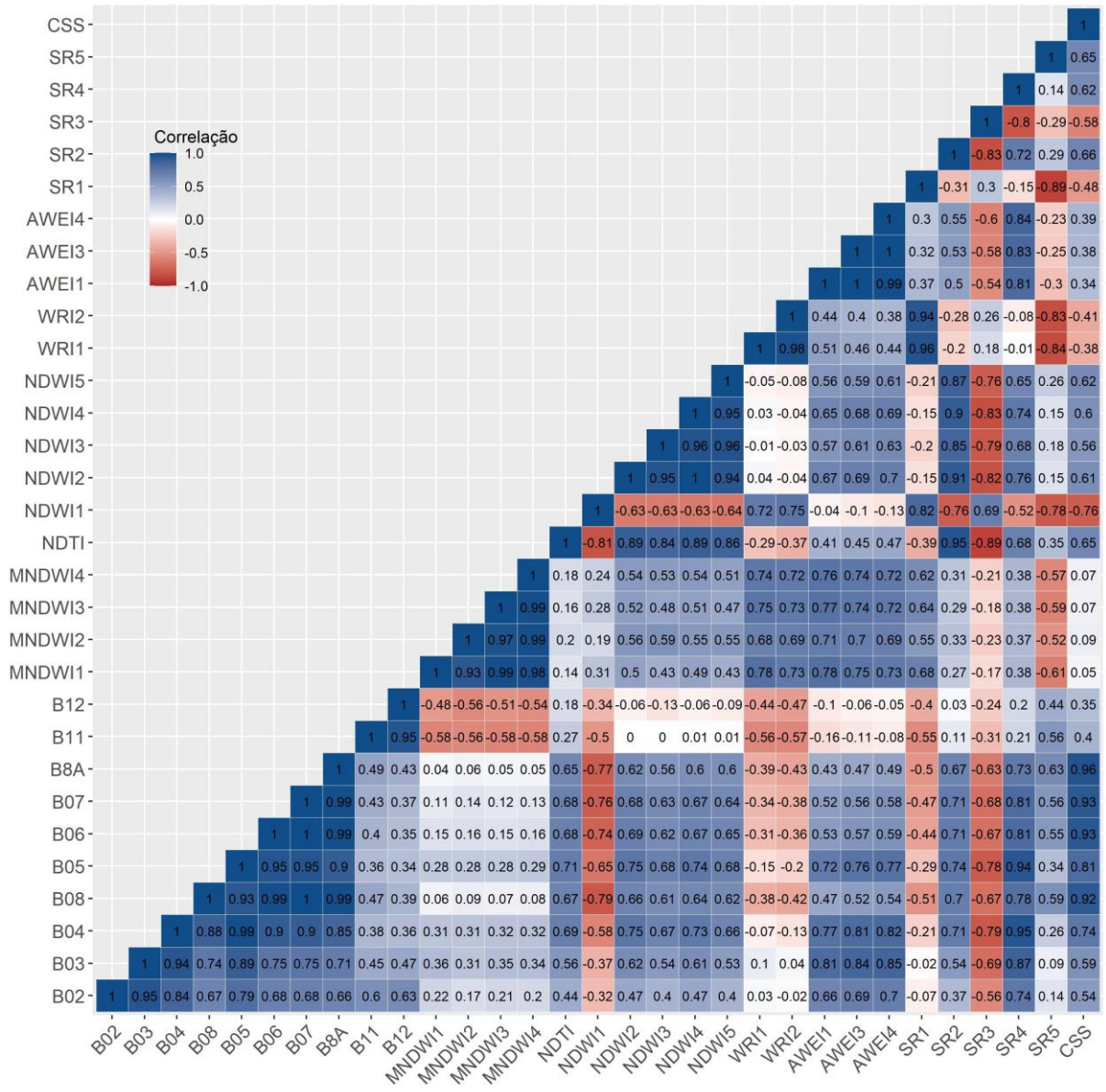
APÊNDICE E. Data da coleta e data da passagem do satélite MSI/Sentinel 2

Estação	Data da coleta	Data da Imagem	Defasagem	Vazão (m³/s)	CSS (mg/L)
56992480	20/12/2016	17/12/2016	3 dias de defasagem	702,00	300,00
56539000	18/02/2017	18/02/2017	Corresponde	97,30	118,90
56992480	29/04/2017	26/04/2017	3 dias de defasagem	191,00	14,00
56992390	29/04/2017	28/07/2017	1 dia de defasagem	219,00	3,00
56992480	04/08/2017	04/08/2017	Corresponde	115,78	16,00
56992390	04/08/2017	04/08/2017	Corresponde	130,77	12,00
RD012	17/10/2017	18/10/2017	1 dia de defasagem	110,53	2,20
RD09	10/03/2018	10/01/2018	2 dias de defasagem	159,53	5,00
RD012	11/01/2018	11/01/2018	Corresponde	558,19	148,01
RD015	19/01/2018	16/01/2018	3 dias de defasagem	281,59	39,00
RD06	22/01/2018	18/01/2018	3 dias de defasagem	14,00	14,00
RD03	17/01/2018	19/01/2018	2 dias de defasagem	71,61	36,00
56110005	25/01/2018	24/01/2018	1 dia de defasagem	26,70	8,10
56994500	03/04/2018	01/04/2018	2 dias de defasagem	557,00	45,90
RD09	05/04/2018	04/04/2018	1 dia de defasagem	427,17	34,00
RD01	12/04/2018	14/04/2018	2 dias de defasagem	100,11	55,17
RD07	09/07/2018	08/07/2018	1 dia de defasagem	186,18	19,00
RD012	18/07/2018	15/07/2018	3 dias de defasagem	226,58	5,00
RD01	19/07/2018	18/07/2018	1 dia de defasagem	46,53	6,00
RD015	20/07/2018	20/07/2018	Corresponde	256,46	14,06
RD012	10/09/2018	08/10/2018	1 dia de defasagem	160,57	6,00
RD06	25/10/2018	26/10/2018	1 dia de defasagem	110,00	110,00
RD012	16/01/2019	16/01/2019	Corresponde	307,33	36,00
RD015	23/01/2019	21/01/2019	2 dias de defasagem	366,91	35,99
RD03	22/01/2019	24/01/2019	2 dias de defasagem	63,69	40,00
56425000	25/01/2019	24/01/2019	1 dia de defasagem	62,80	29,60
56920000	06/02/2019	03/02/2019	3 dias de defasagem	172,00	8,50
56425000	14/03/2019	15/03/2019	1 dia de defasagem	72,60	77,50
56994500	19/03/2019	17/03/2019	2 dias de defasagem	288,00	35,80
RD09	04/04/2019	04/04/2019	Corresponde	239,51	7,00
RD015	24/04/2019	21/04/2019	3 dias de defasagem	459,49	15,00
RD06	25/04/2019	24/04/2019	1 dia de defasagem	38,00	38,00
56539000	17/05/2019	14/05/2019	3 dias de defasagem	173,00	80,20
56425000	20/05/2019	19/05/2019	1 dia de defasagem	96,90	48,50
RD07	02/07/2019	03/07/2019	1 dia de defasagem	151,64	2,10
RD09	04/07/2019	03/07/2019	1 dia de defasagem	156,39	1,30
RD012	15/10/2019	18/10/2019	3 dias de defasagem	142,83	2,10
56539000	12/11/2019	10/11/2019	2 dias de defasagem	82,00	30,50
RD07	07/07/2020	07/07/2020	Corresponde	288,94	17,38
RD03	14/07/2020	12/07/2020	2 dias de defasagem	84,83	14,60
RD012	13/07/2020	14/07/2020	1 dia de defasagem	373,87	8,70

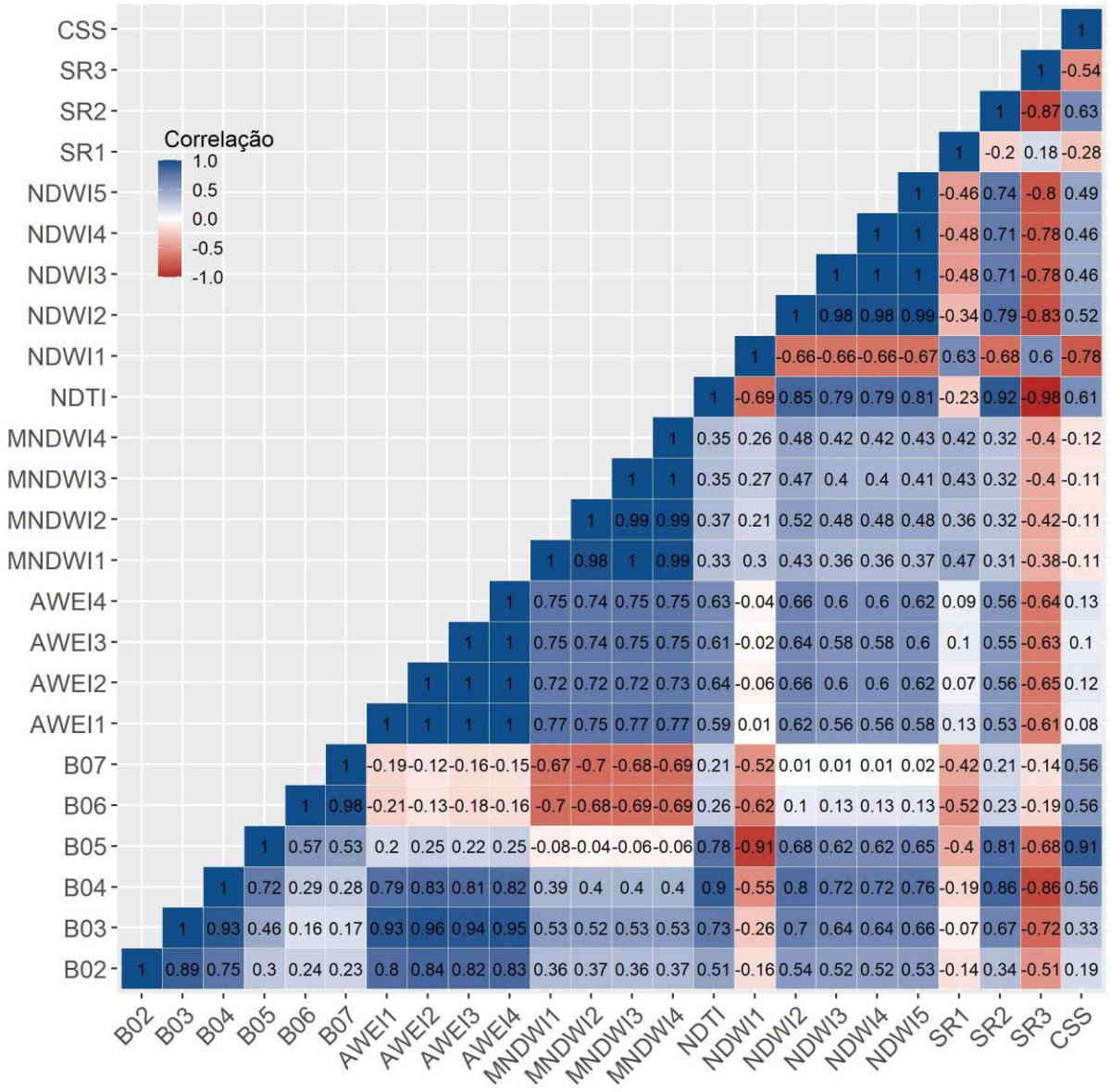
APÊNDICE F. Data da coleta e data da passagem do satélite OLI/Landsat 8

Estação	Data da coleta	Data da Imagem	Defasagem	Vazão (m³/s)	CSS (mg/L)
56425000	17/05/2013	14/05/2013	3 dias de defasagem	101,80	12,30
56920000	28/08/2013	27/08/2013	1 dia de defasagem	224,63	7,00
56992480	10/09/2013	12/09/2013	2 dias de defasagem	205,00	4,96
56992480	19/02/2014	19/02/2014	Corresponde	436,00	17,23
26032015	27/03/2015	26/03/2015	1 dia de defasagem	430,00	19,88
56992390	26/03/2015	26/03/2015	Corresponde	328,00	16,69
56920000	04/08/2015	01/08/2015	3 dias de defasagem	169,00	3,00
56992390	04/11/2015	05/11/2015	1 dia de defasagem	129,00	10,00
56920000	14/12/2015	14/12/2015	Corresponde	519,00	473,00
56920000	17/12/2015	14/12/2015	3 dias de defasagem	374,00	271,70
56920000	19/09/2016	20/09/2016	1 dia de defasagem	78,10	2,80
56539000	18/02/2017	18/02/2017	Corresponde	97,30	118,90
56539000	07/04/2017	07/04/2017	Corresponde	72,60	58,00
56994500	13/04/2017	16/04/2017	3 dias de defasagem	305,00	17,00
56992390	04/08/2017	06/08/2017	2 dias de defasagem	130,77	12,00
RDO-12	11/01/2018	13/01/2018	2 dias de defasagem	558,19	148,01
RDO-06	22/01/2018	20/01/2018	2 dias de defasagem	167,87	14,00
56539000	27/03/2018	25/03/2018	2 dias de defasagem	282,00	300,50
56994500	03/04/2018	03/04/2018	Corresponde	557,00	45,90
RDO-09	05/04/2018	03/04/2018	2 dias de defasagem	427,17	34,00
RDO-15	17/04/2018	19/04/2018	2 dias de defasagem	659,63	46,01
RDO-15	20/07/2018	17/07/2018	3 dias de defasagem	256,46	14,06
56920000	24/09/2018	26/09/2018	2 dias de defasagem	226,00	10,40
RDO-09	03/10/2018	03/10/2018	Corresponde	159,53	5,00
RDO-12	16/01/2019	16/01/2019	Corresponde	307,33	36,00
56425000	25/01/2019	23/01/2019	2 dias de defasagem	62,80	29,60
RDO-03	22/01/2019	23/01/2019	1 dia de defasagem	63,69	40,00
56539000	28/03/2019	28/03/2019	Corresponde	133,00	59,90
RDO-09	04/04/2019	06/04/2019	2 dias de defasagem	239,51	7,00
56920000	25/04/2019	22/04/2019	3 dias de defasagem	354,00	20,50
56539000	17/05/2019	15/05/2019	2 dias de defasagem	173,00	80,20
RDO-09	04/07/2019	02/07/2019	2 dias de defasagem	156,39	1,30
RDO-07	02/07/2019	02/07/2019	Corresponde	151,64	2,10
RDO-15	23/07/2019	20/07/2019	3 dias defasagem	128,22	14,50
RDO-12	15/10/2019	15/10/2019	Corresponde	142,83	2,10
RDO-09	09/01/2020	10/01/2020	1 dia de defasagem	559,83	49,39
RDO-12	21/01/2020	19/01/2020	2 dias de defasagem	636,86	19,30
RDO-12	13/07/2020	13/07/2020	Corresponde	373,87	8,70
RDO-15	22/07/2020	22/07/2020	Corresponde	313,51	18,30

APÊNDICE G. Matriz de correlação entre as variáveis preditoras obtidas a partir do satélite Sentinel 2 entre si e com a concentração superficial de sedimentos (CSS)



APÊNDICE H. Matriz de correlação entre as variáveis preditoras obtidas a partir do satélite Landsat 8 entre si e com a concentração superficial de sedimentos (CSS)



APÊNDICE I. Estações pluviométricas utilizadas para obtenção dos dados observados de precipitação diária na bacia hidrográfica do rio Doce

Lat.	Long.	Cod.	Período	Lat.	Long.	Cod.	Período
-17.846	-42.076	1742017	1976-2020	-19.811	-41.438	1941019	1983-2020
-17.992	-42.394	1742019	1984-2020	-19.416	-41.730	1941021	1995-2020
-18.986	-40.746	1840000	1968-2020	-19.533	-41.009	1941023	2002-2013
-18.575	-41.918	1841001	1940-2020	-19.834	-42.318	1942002	1941-2020
-18.239	-41.749	1841003	1940-2020	-19.999	-42.348	1942006	1946-2020
-18.976	-41.640	1841011	1974-2020	-19.374	-42.105	1942008	1969-2020
-18.850	-41.933	1841015	1984-2020	-19.525	-42.644	1942029	1986-2020
-18.777	-41.483	1841019	1984-2020	-19.316	-42.396	1942030	1986-2020
-18.883	-41.950	1841020	1985-2020	-19.777	-42.477	1942031	1986-2020
-18.363	-42.602	1842004	1940-2020	-19.189	-42.423	1942032	1986-2020
-18.612	-42.279	1842005	1941-2020	-19.873	-42.132	1942048	2006-2020
-18.772	-42.931	1842007	1946-2020	-19.923	-43.178	1943001	1940-2020
-18.201	-42.455	1842008	1969-2020	-19.017	-43.444	1943002	1940-2019
-18.553	-42.764	1842020	1984-2020	-19.250	-43.014	1943003	1940-2020
-18.280	-43.001	1843012	1984-2020	-19.945	-43.401	1943007	1940-2019
-18.593	-43.413	1843011	1984-2020	-19.440	-43.119	1943008	1940-2020
-19.578	-39.794	1939002	1974-2020	-19.218	-43.374	1943025	1945-2020
-19.874	-40.874	1940000	1944-2020	-19.881	-43.368	1943027	1946-2020
-19.805	-40.679	1940001	1947-2020	-19.767	-43.026	1943100	2003-2020
-19.692	-40.398	1940005	1948-2020	-20.108	-41.728	2041008	1946-2020
-19.531	-40.623	1940006	1967-2020	-20.079	-41.121	2041023	1967-2020
-19.220	-40.853	1940009	1957-2020	-20.171	-41.961	2041048	1983-2020
-19.664	-40.835	1940012	1957-2020	-20.104	-42.440	2042008	1941-2020
-19.238	-40.591	1940013	1969-2020	-20.299	-42.478	2042010	1942-2020
-19.058	-40.516	1940016	1968-2020	-20.726	-42.917	2042015	1960-2008
-19.955	-40.742	1940020	1970-2020	-20.683	-42.807	2042016	1967-2020
-19.274	-40.321	1940023	1970-2020	-20.277	-42.326	2042017	1967-2020
-19.295	-40.518	1940025	1970-2010	-20.385	-42.903	2042018	1975-2020
-19.508	-40.864	1940047	2003-2013	-20.011	-42.674	2042031	1981-2020
-19.799	-41.706	1941000	1941-2018	-20.714	-43.000	2042040	2008-2019
-19.525	-41.015	1941003	1941-2020	-20.363	-43.144	2043009	1941-2020
-19.343	-41.246	1941004	1941-2019	-20.691	-43.299	2043010	1941-2020
-19.062	-41.533	1941005	1944-2020	-20.390	-43.180	2043011	1941-2020
-19.595	-41.458	1941006	1946-2019	-20.670	-43.088	2043014	1941-2020
-19.901	-41.058	1941008	1947-2020	-20.517	-43.017	2043025	1959-2020
-19.691	-41.020	1941009	1967-2020	-20.848	-43.242	2043026	1967-2019
-19.493	-41.162	1941010	1967-2020	-20.286	-43.099	2043027	1967-2020
-19.678	-41.836	1941011	1970-2020	-20.097	-43.488	2043059	1983-2020
-19.059	-41.028	1941012	1970-2019	-21.149	-43.520	2143003	1940-2020
-19.162	-41.862	1941018	1984-2020				

APÊNDICE J. Mudanças nas classes de uso e cobertura da terra entre 1985 a 2019 na região alto rio Doce da bacia hidrográfica do rio Doce

Ano	Corpos de água (km²)	Área não vegetada (km²)	Formação não florestal (km²)	Mosaico Agricultura e Pastagem (km²)	Agricultura (km²)	Floresta plantada (km²)	Floresta nativa (km²)	Pastagem (km²	Mineração (km²)	Afloramento rochoso (km²)
1985	25,8	54,0	451,9	945,3	1236,7	1251,2	3140,0	8710,7	20,9	21,7
1986	26,7	51,4	430,5	1012,5	1315,0	1261,4	3108,0	8603,9	21,2	26,2
1987	27,7	54,4	404,4	1011,3	1315,2	1240,9	3036,0	8712,7	21,6	30,8
1988	27,8	56,2	400,2	940,1	1267,5	1222,2	3045,6	8841,0	21,8	32,2
1989	27,8	56,6	428,0	825,3	1138,1	1182,0	3092,2	9052,7	21,8	30,8
1990	27,8	56,5	419,8	831,6	1138,2	1167,6	3085,1	9078,4	22,0	29,1
1991	27,9	56,5	417,8	819,7	1132,3	1180,5	3023,9	9145,7	22,2	29,4
1992	28,2	55,7	426,2	876,7	1147,9	1175,2	3017,0	9078,2	22,1	29,1
1993	28,1	57,5	422,6	840,4	1132,0	1165,1	3107,0	9049,9	21,9	31,1
1994	27,2	60,1	400,2	914,1	1234,9	1206,9	3115,6	8839,8	21,9	33,6
1995	27,2	62,4	402,4	955,7	1293,3	1225,9	3129,7	8702,5	22,3	33,0
1996	27,1	66,2	405,7	939,5	1284,8	1225,9	3133,4	8714,7	22,4	34,0
1997	27,3	67,8	395,3	949,7	1283,1	1219,7	3138,9	8715,1	22,6	34,5
1998	27,5	67,0	379,5	988,0	1308,8	1202,2	3107,1	8714,7	23,0	35,6
1999	29,7	70,2	350,5	1054,5	1340,6	1166,6	3004,2	8775,1	23,7	37,9
2000	29,8	73,9	384,1	886,7	1197,6	1166,2	3081,8	8967,4	24,0	40,8
2001	30,2	75,9	396,5	837,1	1174,1	1181,6	3112,3	8977,5	24,2	42,4
2002	30,4	76,9	407,5	794,3	1103,3	1161,0	3106,9	9105,5	24,4	41,7
2003	31,7	74,4	435,8	671,6	984,2	1146,8	3152,6	9289,6	24,7	41,0
2004	32,4	76,5	460,4	627,1	934,4	1164,0	3184,7	9306,8	24,6	40,9
2005	33,4	79,6	431,0	725,1	1055,6	1184,9	3092,4	9185,2	24,8	40,0
2006	33,7	80,9	425,5	696,0	1058,9	1208,6	3054,2	9231,2	24,8	39,4
2007	33,7	82,4	451,4	654,3	1034,0	1250,2	3186,6	9093,9	24,8	41,4
2008	33,8	84,5	434,5	757,4	1155,1	1304,6	3190,3	8824,3	24,7	42,8
2009	33,8	90,3	416,4	869,9	1279,3	1366,3	3222,9	8501,6	24,7	45,2

Continua...

Apêndice J. Continuação

2010	34,2	94,5	430,2	898,9	1288,3	1406,8	3307,4	8318,8	24,3	46,5
2011	34,2	93,7	441,8	925,1	1334,8	1467,3	3406,1	8076,6	24,2	46,6
2012	34,3	95,6	448,1	980,4	1398,2	1516,5	3450,0	7855,5	24,5	46,6
2013	34,3	99,4	446,4	1009,6	1419,4	1521,6	3444,2	7802,4	24,3	47,3
2014	34,3	99,3	452,4	964,6	1373,5	1513,3	3465,4	7874,6	24,1	47,6
2015	34,4	101,2	447,2	988,3	1406,9	1513,3	3442,9	7840,5	24,6	49,0
2016	33,6	104,8	436,2	1099,0	1474,9	1552,2	3404,6	7663,8	27,0	51,2
2017	33,3	105,7	434,3	1159,4	1557,7	1575,7	3379,7	7521,0	27,1	53,3
2018	33,7	111,0	447,1	1226,7	1625,9	1602,5	3375,7	7344,9	27,3	51,0
2019	33,6	111,2	452,8	1298,7	1655,0	1590,9	3291,6	7341,1	27,1	45,9

APÊNDICE K. Mudanças nas classes de uso e cobertura da terra entre 1985 a 2019 na região baixo rio Doce da bacia hidrográfica do rio Doce

Ano	Corpos de água (km²)	Área não vegetada (km²)	Formação não florestal (km²)	Mosaico Agricultura e Pastagem (km²)	Agricultura (km²)	Floresta plantada (km²)	Floresta nativa (km²)	Pastagem (km²	Mineração (km²)	Afloramento rochoso (km²)
1985	186,1	315,3	2393,3	3845,0	5339,2	6178,2	20226,4	37704,4	110,8	188,4
1986	188,3	321,2	2285,1	3919,8	5477,0	6158,5	20059,3	37750,3	111,2	208,3
1987	189,0	327,0	2246,1	3921,2	5566,4	6236,1	20051,1	37595,6	112,6	225,1
1988	189,0	338,7	2281,0	3638,1	5353,9	6231,5	20231,0	37861,1	113,3	231,4
1989	188,8	342,8	2281,4	3517,8	5272,1	6200,5	20401,6	37918,6	113,4	231,5
1990	187,9	342,5	2287,1	3370,7	5107,4	6134,6	20355,1	38338,1	113,8	232,3
1991	189,6	343,3	2366,3	3152,3	4845,4	6247,5	20173,8	38807,5	115,4	229,6
1992	190,4	336,2	2402,1	3243,3	4754,2	6203,2	20115,2	38886,5	116,0	225,7
1993	192,2	344,8	2342,9	3144,1	4724,5	6110,2	20024,2	39245,5	114,4	230,4
1994	190,3	353,7	2310,6	3376,7	4912,8	6256,7	19836,0	38882,5	115,0	235,7
1995	189,6	367,4	2284,3	3586,1	5181,2	6354,4	19781,2	38381,1	115,9	231,2
1996	189,3	378,9	2245,6	3653,8	5282,5	6377,9	19857,1	38131,5	115,8	235,9
1997	188,9	386,7	2200,9	3796,0	5425,2	6415,9	19827,7	37869,9	117,3	239,1
1998	189,4	387,8	2188,1	3855,1	5447,3	6345,0	19739,8	37952,4	118,8	241,6
1999	190,8	408,4	2108,2	3897,9	5395,2	6199,4	19297,9	38589,7	120,9	255,0
2000	192,7	418,5	2154,8	3576,7	4978,6	6083,0	19168,8	39499,8	121,8	264,7
2001	193,2	429,3	2120,6	3580,3	5104,2	6161,9	19052,4	39418,1	122,5	273,3
2002	195,5	437,9	2166,2	3539,1	4953,7	6131,7	19027,0	39606,3	123,9	275,9
2003	197,3	438,9	2213,5	3389,7	4755,8	6127,6	19090,1	39840,8	124,4	277,2
2004	198,9	449,8	2202,8	3451,7	4949,0	6266,5	18923,5	39615,1	124,9	274,5
2005	203,5	456,1	2096,4	3906,5	5546,5	6418,7	18704,9	38725,5	126,0	272,8
2006	214,6	459,9	2103,6	3863,3	5567,5	6538,4	18873,8	38438,7	126,5	272,8
2007	214,4	470,1	2163,7	3856,8	5498,1	6666,5	19253,4	37923,7	126,7	283,3
2008	215,5	471,3	2130,7	4166,6	5752,3	6840,3	19382,3	37082,9	126,4	286,9
2009	218,8	490,8	2085,7	4223,0	5925,3	7038,3	19567,1	36488,6	126,6	288,9

Continua...

Apêndice K. Continuação

2010	219,4	504,3	2118,0	4190,7	5857,3	7158,9	19744,9	36241,5	125,5	291,2
2011	220,5	510,0	2170,6	4094,2	5752,8	7266,4	20066,2	35954,5	127,3	290,3
2012	220,8	511,6	2191,2	4145,5	5740,0	7336,0	20422,6	35463,5	130,0	289,8
2013	219,7	524,9	2180,2	4304,0	5788,4	7349,1	20446,9	35219,5	130,5	287,4
2014	219,0	532,0	2186,6	4384,6	5717,5	7330,8	20622,2	35033,6	132,2	290,8
2015	218,5	538,7	2231,1	4484,1	5631,3	7281,1	20887,5	34740,6	134,2	298,5
2016	216,6	549,9	2147,7	5098,7	6034,5	7467,8	20911,9	33569,6	146,8	301,3
2017	213,7	558,5	2180,3	5234,1	6228,6	7617,3	21122,1	32833,5	146,3	308,8
2018	217,2	578,3	2193,5	5843,1	6583,0	7768,1	21410,9	31389,4	147,5	307,1
2019	215,4	594,1	2187,8	6397,3	6735,0	7729,2	21045,1	31103,9	145,3	292,1

ANEXOS

ANEXO A. Informações técnicas das bandas da constelação do satélite MSI/Sentinel 2

Quantização		12 bits	
Resolução Temporal		[1]5 dias	
Banda	**Descrição**	**Comprimento de onda (nm)**	**Resolução (m)**
B02	Azul	439 - 535	10
B03	Verde	537 - 582	10
B04	Vermelho	646 - 685	10
B08	Infravermelho Próximo (NIR)	767 - 908	10
B05	Visível e infravermelho próximo (VNIR 1)	694 - 714	20
B06	Visível e infravermelho próximo (VNIR 2)	731 - 749	20
B07	Visível e infravermelho próximo (VNIR 3)	768 - 796	20
B08A	Visível e infravermelho próximo (VNIR 4)	848 - 881	20
B11	Infravermelho de ondas curtas (SWIR 1)	1539 - 168	20
B12	Infravermelho de ondas curtas (SWIR 2)	2072 - 2312	20
B01	Aerossol	421 - 457	60
B09	Vapor de água	931 - 958	60
B10	Cirrus	1338 - 1414	60

[1]A partir de março de 2017, após o lançamento do satélite Sentinel 2B

Fonte: ESA (2018)

ANEXO B. Informações técnicas das bandas do satélite OLI/Landsat 8

Quantização		12 bits	
Resolução temporal		16 dias	
Bandas	**Descrição**	**Comprimento de onda (nm)**	**Resolução (m)**
B01	Aerossol	433 - 453	30 m
B02	Azul	452 - 512	30 m
B03	Verde	533 - 590	30 m
B04	Vermelho	636 - 673	30 m
B05	Infravermelho próximo (NIR)	851 - 879	30 m
B06	Infravermelho de ondas curtas (SWIR 1)	1567 - 1651	30 m
B07	Infravermelho de ondas curtas (SWIR 2)	2107 - 2294	30 m
B08	Pancromático	500 - 680	15 m
B09	Cirrus	1360 - 1380	30 m
B10	Infravermelho termal (TIR - 1)	10300 - 11300	100 m
B11	Infravermelho termal (TIR - 2)	11500 - 12510	100 m

Fonte: NASA (2018b)