



**UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA CIVIL
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL**

LUIS GUSTAVO DE MOURA REIS

**APRENDIZADO DE MÁQUINA NA PREVISÃO DE VAZÕES MÍNIMAS A
PARTIR DE ÍNDICES DE SECA E PRODUTOS DERIVADOS DE
SENSORIAMENTO REMOTO EM ESCALA GLOBAL**

Recife
2022

LUIS GUSTAVO DE MOURA REIS

**APRENDIZADO DE MÁQUINA NA PREVISÃO DE VAZÕES MÍNIMAS A
PARTIR DE ÍNDICES DE SECA E PRODUTOS DERIVADOS DE
SENSORIAMENTO REMOTO EM ESCALA GLOBAL**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal de Pernambuco como parte dos requisitos exigidos para obtenção do título de Doutor em Engenharia Civil.

Área de Concentração: Tecnologia Ambiental e Recursos Hídricos.

Orientadora: Profa. Dra. Suzana Maria Gico Lima Montenegro - UFPE

Recife
2022

Catálogo na fonte
Bibliotecário Gabriel Luz, CRB-4 / 2222

R375a Reis, Luis Gustavo de Moura.
Aprendizado de máquina na previsão de vazões mínimas a partir de índices de seca e produtos derivados de sensoriamento remoto em escala global / Luis Gustavo de Moura Reis. 2022.
134 f: il.

Orientadora: Profa. Dra. Suzana Maria Gico Lima Montenegro.
Tese (Doutorado) – Universidade Federal de Pernambuco. CTG.
Programa de Pós-Graduação em Engenharia Civil, Recife, 2022.
Inclui referências e apêndice.

1. Engenharia civil. 2. Previsão de longo-prazo. 3. Previsão de vazões mínimas. 4. Aprendizado de máquina. 5. Aprendizado estatístico. 6. SVM; 7. RF. 8. XGBoost. I. Montenegro, Suzana Maria Gico Lima (Orientadora). II. Título.

UFPE

624 CDD (22. ed.)

BCTG / 2023 - 23

LUIS GUSTAVO DE MOURA REIS

**APRENDIZADO DE MÁQUINA NA PREVISÃO DE VAZÕES MÍNIMAS
A PARTIR DE ÍNDICES DE SECA E PRODUTOS DERIVADOS DE
SENSORIAMENTO REMOTO EM ESCALA GLOBAL**

Tese apresentada ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, como requisito para obtenção do título de Doutor em Engenharia Civil, Área de Concentração Tecnologia Ambiental e Recursos Hídricos.

Aprovada em 27/07/2022

Orientadora: Prof.^a Dr.^a Suzana Maria Gico Lima Montenegro - UFPE

BANCA EXAMINADORA

participação por videoconferência

Prof. Dr. André Ricardo Backes (examinador externo)
Universidade Federal de Uberlândia

participação por videoconferência

Prof. Dr. Benedito Cláudio da Silva (examinador externo)
Universidade Federal de Itajubá

participação por videoconferência

Prof. Dr. Adriano Rolim da Paz (examinador externo)
Universidade Federal da Paraíba

participação por videoconferência

Prof.^a Dr.^a Dóris Regina Alves Veleda (examinadora externa)
Universidade Federal de Pernambuco

participação por videoconferência

Prof. Dr. Alfredo Ribeiro Neto (examinador interno)
Universidade Federal de Pernambuco

On the artist's journey, we don't get better by increments, we get better by fits and starts.

The trajectory is not a smoothly-ascending curve, but a herky-jerky spasm-fest marked by seeming dead-ends, plateaus, dark nights of the soul, intervals of boredom and stasis, not to mention bouts of terror, despair and self-doubt, which are followed, if we're lucky, by quantum leaps to the next level.

In other words, we advance by breakthroughs.

(Steven Pressfield, 2012)

*Gratidão a Deus, pela sua graça e
infinita misericórdia.*

AGRADECIMENTOS

A Deus por sua graça, por seus dons e seu sustento e principalmente por ter me permitido essa experiência grandiosa. À minha esposa Rys, companheira de jornada, mulher de fé, que me incentivou a buscar esse desafio e a persistir diante de tantas dificuldades. À minha mãe Verônica e meu pai Luis Carlos que nos deixou durante a caminhada, pela minha vida e pela educação me dada. À minha avó, professora Olga, pelas longas aulas matinais na velha mesa de madeira no quintal, zelando sempre pelos compromissos escolares dos seus netos. Aos meus irmãos Felipe e Ana, pelo carinho de sempre. A minha sobrinha Luisa pela sua alegria e criatividade sempre contagiante. Ao professor Ranilson França (*in memorian*) por todo seu apoio, sua sabedoria e sua cultura. Aos meus gatos Cheetaro, Nininha, Lila Ratuska (liloca) e Sussu, pela companhia nas varadas de noite em longos estudos e produção intelectual. Aos meus mentores Dr. Alberto Schwartzman, Prof. Valmir Pedrosa e Prof. Roberaldo Carvalho, pela orientação, mentoria e tutoria contínua durante todos esses anos. Aos professores e colegas do GRH da UFPE por toda dedicação, apoio e direção. À secretaria da Pós-Civil, nas pessoas da Andrea Negromonte e Cleide, sempre zelando com cuidado pela nossa vida acadêmica e sempre de prontidão para nossas urgências. Ao departamento de oceanografia da UFPE e ao Instituto de Ciências Atmosféricas da UFAL por todo o conhecimento transmitido a esse engenheiro civil. Ao grande parceiro de caminhada, disciplinas e produção acadêmica, Dr. Wendson Souza por todo suporte mútuo. Ao prof. Antonio Miguel Ruiz Armenteros por todo apoio e a universidade de Jaén-ES, por todo apoio e supervisão em nossas publicações. À editora MDPI pela sua agilidade e profissionalismo. Ao meu professor de Inglês André Batista, a quem sou grato por poder transmitir minha mensagem ao mundo científico. Ao MIT por disponibilizar suas aulas de inteligência artificial do prof. Patrick Winston (*in memorian*) acessíveis e abertas a todo mundo. Aos professores Aboukadel Kassambara e Jason Brownlee, pelo conteúdo disponibilizado e publicações na área de Machine Learning. A toda comunidade da linguagem R, sempre disponibilizando novas ferramentas e bibliotecas gratuitas, sem as quais seria impossível realizar nossos experimentos. Aos sites e blogs *StackOverflow*, *Stathmethods*, *RBloggers*, *AnalyticsVidhya*, *TowardsDataScience* e outros, mantidos pela comunidade incansável de IA, sempre nos socorrendo com uma solução quando temos problemas – nunca estamos sós. Aos cientistas Grey S. Nearing e Frederick Kratzert que encabeçaram uma provocação sem precedentes na comunidade científica, e foi o que me motivou a concluir esse trabalho. À obra do cientista Vladimir Vapnik. À obra de Steven Pressfield, Tim

Ferrys, Brené Brown, Seth Godin, Victoria Labalme, que me ajudou a nos manter escrevendo e criando, mesmo quando não tinha vontade. Aos meus jovens colegas: Marina Moura, Artur Brasil, Rafael Matias e Wallison Carvalho – por toda inspiração e motivação. Ao engenheiro e cientista Denis Duda Costa, pela sua inspiração, raciocínio lógico, que muito me ajudou a sair de encruzilhadas. A colega e amiga Maria Elisa Costa, por seu apoio, cobranças e pela sua prontidão em ajudar. Ao Padre Marcio Roberto e Padre Marcos, pela direção na nossa batalha espiritual diária na perseverança da nossa missão.

À Professora Suzana Montenegro e ao Professor Carlos Ruberto, por aceitarem me orientar nessa jornada longa, pelo irrestrito apoio que foi me dado, mesmo diante de tantas quebras de compromisso, sempre acreditaram que eu poderia chegar até aqui. Sempre com palavras positivas, julgando menos e apoiando mais. Isso para mim foi fundamental.

A “minha” baleia Moby Dick, que me proporcionou essa luta justa. Adversário mortal, que por vezes quase me venceu, destruiu meu barco, me deixou à deriva. Mas fui eu quem a perseguiu primeiro. Onde e quando nos encontraremos novamente?

RESUMO

A previsão de longo prazo de vazões mínimas é de grande relevância num cenário provocado mudanças nos padrões de precipitação, ocorrência de secas mais severas e prolongadas. Num cenário de mudanças de tendências, medidas regulatórias de controle baseadas em vazões de referência estáticas ou baseadas na climatologia das vazões serão ineficazes, podendo ser permissivas em períodos críticos e restritivas em períodos favoráveis. De um lado, é crescente a disponibilização de dados em escala global e produtos derivados de sensoriamento remoto com alta resolução temporal e espacial. De outro, os modelos gerados por algoritmos de aprendizado estatístico (*machine learning*) têm se mostrado cada vez mais acurados e robustos no processo de aprendizado e previsão. Entretanto, seu uso ainda é incipiente no campo das ciências hidrológicas. Neste trabalho foram utilizados três algoritmos de aprendizado de máquina: *Support Vector Machine* (SMV), *Random Forest* (RF) e *Extreme Gradient Boosting* (XGBoost) para realizar previsões de vazões mínimas com 6-9 meses de antecedência, em cinco bacias localizadas no Cerrado e Mata-Atlântica. Como preditores foram utilizados os índices *Temperature Condition Index* (TCI), *Vegetation Condition Index* (VCI), *Palmer meteorológico* (PDSI), vazões observadas (Q) e precipitação obtida pelo *Climate hazards infrared precipitation with stations* (CHIRPS). Foram investigadas as performances de aprendizado dos algoritmos no período de extrema seca não utilizada no treinamento, o grau de importância dada aos preditores, a influência da variabilidade espacial na performance e a capacidade de generalização. A habilidade do índice Palmer como substituto da vazão observada também foi investigada. Os modelos obtidos pelo algoritmo SVM apresentaram desempenho drasticamente superior aos demais, com grande poder de generalização, especialmente quando aplicados de forma cruzada em bacias localizadas em biomas, região climática, altura pluviométrica e capacidade de recarga distintas. Os resultados mostraram que a previsibilidade de vazões mínimas em locais com baixos índices pluviométricos e precipitação irregular é controlada fortemente por uma componente interanual. Através da vazão observada, associada a técnicas de suavização na escala temporal adequada e diferenciação do sinal, o algoritmo SVM conseguiu realizar o aprendizado sobre o balanço hidrológico, sendo essa variável fundamental no processo de previsão.

Palavras-chave: previsão de longo-prazo; previsão de vazões mínimas; aprendizado de máquina; aprendizado estatístico; SVM; RF; XGBoost.

ABSTRACT

Long-term low-flow predictions are of great importance in a scenario marked by changes in precipitation patterns and increasing drought frequency, severity, and duration. In the presence of significative hydrological trends, water permit limitations based on permanent low flow quantiles may not work out, leading to incompatible restrictions during the wet or dry season. On one side is the increasing availability of climatic data on a global scale and satellite-derived data, with high spatial and temporal resolution. On another, data-driven models derived from machine learning algorithms have shown powerful skill in dealing with loads of data and delivering accurate results in prediction tasks. Nevertheless, its application in hydrology sciences is still underused. This work employed machine learning algorithms associated with satellite-derived and reanalysis data to forecast low streamflows in the dry season with 6- and 9-months lead time. Global drought indexes: Temperature Condition Index - TCI and Vegetation Condition Index - VCI provided from NOAA/AVHRR-VIIRS, and Palmer Severity Drouth Index – PDSI were employed as predictors. Climate hazards infrared precipitation with stations – CHIRPS was considered as input. We also investigate the skill PDSI as a proxy to observed streamflows. The skill of machine learning algorithms: SVM (support vector machine), RF (random forest), and XGBoost (Extreme Gradient Boosting) were also investigated. The experiment was carried out in five watersheds located in Brazilian Atlantic Forest and Cerrado Biome (Brazilian Savana). Algorithms' performance, generalization ability, and skill of predictors were tested during the most severe drought period registered. Influences of spatial variability were investigated by the extraction of covariation modes and EOFs. Our experiments showed that an inter-annual low-frequency signal controls the predictability of low streamflows. The lower the annual rainfall, the higher interannual variability. Machine Learning models derived from the SVM-Linear algorithm drastically outperformed RF and XGBoost and showed great skill in predicting low flows with a six-month lead time. SVM showed great generalization skills when the models were submitted to a proxy basin test. Observed streamflows play an essential role in the prediction, when associated with smoothing and differentiation technics can drastically improve algorithm's learning about hydrologic water balance.

Keywords: TCI; VCI; PDSI; CHIRPS; seasonal forecasting; low-flow forecast

LISTA DE ILUSTRAÇÕES

- Figura 1- Fluxo de um modelo típico de Machine Learning. A partir de dados e observações históricas se constrói um modelo utilizando um algoritmo ML. Em seguida se faz a avaliação de performance e se otimiza a acurácia e escalabilidade para se ajustar aos requisitos da modelagem. Com o modelo ajustado se realiza a previsão sobre novos dados..... 28
- Figura 2- AI tradicional vs. Machine Learning AI. Nas aplicações de AI tradicional se fornece ao computador um conjunto de regras. Quando se fornecem os dados, ele fornece as respostas relevantes. Na abordagem Machine Learning se fornecem os dados ao computador e as respostas, então o computador aprende as regras por ele mesmo. Quando se fornecem novos dados ele utiliza essas regras para fornecer as novas respostas..... 29
- Figura 3- Funcionamento dos algoritmos supervisionados versus não-supervisionados. Os supervisionados constroem um modelo a partir de dados classificados. Os não-supervisionados extraem padrões a partir de dados não-classificados..... 31
- Figura 4- Análise comparativa entre o erro em uma amostra de treino e a amostra de teste variando-se o grau de complexidade do modelo. As linhas azuis se referem aos erros na amostra de treino e as linhas vermelhas se referem ao erro na amostra de teste. Experimento realizado em 100 amostras de treino de tamanho 50 cada. As linhas sólidas ilustram os erros esperados..... 34
- Figura 5- Princípio de minimização do risco estrutural..... 36
- Figura 6- Ilustração do processo de otimização em que o algoritmo SVM procura um hiperplano (linha sólida) que atravessa o espaço de características. O hiperplano ótimo é o que maximiza a margem (linhas tracejadas). A margem é a região no entorno do hiperplano que toca o mínimo de elementos possível, esses elementos que definem a margem são chamados de vetores de suporte ilustrados com contorno nos círculos..... 37
- Figura 7- O algoritmo SVM mapeia os dados de entrada vindo do espaço original num espaço de maiores dimensões chamado espaço de características..... 37
- Figura 8- Ilustração da aplicação de um Kernel onde o algoritmo SVM adiciona uma dimensão a mais para realizar a separação dos dados não linearmente separáveis em seu espaço original. A dimensão adicional no espaço de características permite com que os dados sejam separados por um hiperplano. Quando os dados são projetados novamente em duas dimensões o hiperplano aparece como um limite de decisão curva..... 38

Figura 9- Ilustração de exemplos de aplicação de vários tipos de funções Kernel. Para cada exemplo, a linha sólida indica o limite de decisão reprojetoado no espaço de características original e as linhas tracejadas indicam as margens.....	39
Figura 10-Ilustração do mecanismo de funcionamento do algoritmo random forest (RF)	42
Figura 11- Ilustração do mecanismo de funcionamento do algoritmo Adaptive boosting with decision trees (RF).....	44
Figura 12-Ilustração de funcionamento de uma PC em um espaço tridimensional.....	49
Figura 13-Fluxograma geral da metodologia.....	58
Figura 14-Localização da área de estudo (Elaborado a partir de informações disponibilizadas no SNIRH/ANA, IBGE e MMA).....	60
Figura 15-Isoietas de precipitação anual elaborada com dados no período de 1977-2006 (CPRM; ANA, 2016) produzidos a partir de dados da rede hidrometeorológica nacional e classificação climática de Köppen elaborado a partir de.....	60
Figura 16-Recarga anual e vulnerabilidade das reservas subterrâneas.....	60
Figura 17-Climatologia de precipitações médias sobre as bacias selecionadas (mm·mês ⁻¹) elaborado a partir do produto CHIRPS período (1981-2020).....	61
Figura 18-Climatologia de vazões (mm·mês ⁻¹) elaborado a partir dos dados disponibilizados pela ANA no período de 1980-2020.....	61
Figura 19-Variabilidade temporal índice Palmer (PDSI) em valores médios mensais por bacia no período de 1981-2020. Destaque em vermelho para o período de validação.....	72
Figura 20-Metodologia genérica de aplicação do método K-folds (k=5) na calibração dos modelos estatísticos.....	74
Figura 21-Tendências no período de 1981-2020 das variáveis precipitação (coluna 1), vazão média no período chuvoso ONDJFM (coluna 2), vazão mínima em setembro (coluna 3), índice Palmer PDSI (coluna 4), índice de temperatura TCI (coluna 5) e índice de vegetação VCI (coluna 6). Dados por bacia: entrada em linhas.....	76
Figura 22-Matriz de correlações cruzadas entre as variáveis: POND, PONDJFM, QOND, QONDJFM, PDSIOND, PDSIONDJFM, TCIOND, TCIONDJFM, VCIOND, VCIONDJFM, QSET.....	78
Figura 23-Transformações Wavelets-Coherence (WTC) entre os sinais do índice Palmer (PDSI) e as vazões observadas.....	79
Figura 24-Transformações Continuous Wavelet Transform (CWT) para a série de vazões mínimas QSET* no mês de setembro, escala temporal em anos.....	80
Figura 25-Variáveis adimensionais QSET*, POND*, PONDJFM*, QOND*, QONDJFM*, PDSIOND, VCIOND, TCIOND.....	81

Figura 26-Previsão de vazão Exp2 (QSET*) com 9 meses de antecedência bacia ALTO_JQ com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: POND*, QOND*, PDSIOND, VCIOND, TCIOND.....	84
Figura 27-Previsão de vazão Exp2 (QSET*) com 9 meses de antecedência bacia ALTO_SF com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: POND*, QOND*, PDSIOND, VCIOND, TCIOND.....	84
Figura 28-Previsão de vazão Exp2 (QSET*) com 9 meses de antecedência bacia VACARIA com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: POND*, QOND*, PDSIOND, VCIOND, TCIOND.....	85
Figura 29-Previsão de vazão Exp2 (QSET*) com 9 meses de antecedência bacia Rib_SF com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM Radial), XGBoost (XGB). Preditores utilizados: POND*, QOND*, PDSIOND, VCIOND, TCIOND.....	85
Figura 30-Previsão de vazão Exp2 (QSET*) com 9 meses de antecedência bacia GONGOGI com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: POND*, QOND*, PDSIOND, VCIOND, TCIOND.....	86
Figura 31-Previsão de vazão Exp2 (QSET*) com 6 meses de antecedência bacia ALTO_SF com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). . Preditores utilizados: PONDJFM*, QONDJFM*, PDSIONDJFM, VCIONDJFM, TCIONDJFM	87
Figura 32-Previsão de vazão Exp2 (QSET*) com 6 meses de antecedência bacia ALTO_JQ com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: PONDJFM*, QONDJFM*, PDSIONDJFM, VCIONDJFM, TCIONDJFM	87
Figura 33-Previsão de vazão Exp2 (QSET*) com 6 meses de antecedência bacia VACARIA com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: PONDJFM*, QONDJFM*, PDSIONDJFM, VCIONDJFM, TCIONDJFM	88
Figura 34-Previsão de vazão Exp2 (QSET*) com 6 meses de antecedência bacia Rib_SF com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: PONDJFM*, QONDJFM*, PDSIONDJFM, VCIONDJFM, TCIONDJFM.....	88
Figura 35-Previsão de vazão Exp2 (QSET*) com 6 meses de antecedência bacia GONGOGI com os modelos ML: linear (lm), random forest (RF), máquina de	

vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: PONDJFM*, QONDJFM*, PDSIONDJFM, VCIONDJFM, TCIONDJFM	89
Figura 36-Previsão de vazão Exp3 (QSET*) com 6 meses de antecedência bacia do ALTO_SF. Os preditores são as EOF's dos campos de precipitação CHIRPS, TCI, VCI, PDSI. Previsões com uso de modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).....	91
Figura 37- Padrões modais das funções empíricas ortogonais EOFs da precipitação (CHIRPS), TCI, VCI e índice Palmer (PDSI) para a bacia do Alto São Francisco ALTO_SF.....	92
Figura 38-Autocorrelação (ACF) e autocorrelação parcial (PACF) das séries temporais anuais de QSET* e PDSIONDJFM.....	95
Figura 39-Previsão de vazão Exp4 (QSET*) com 6 meses de antecedência bacia do ALTO_SF Os preditores são TCI PONDJFM*, VCI PONDJFM*, PDSI PONDJFM*, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_SMA, QSET*_SMA.3.diff. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).....	97
Figura 40-Previsão de vazão Exp4 (QSET*) com 6 meses de antecedência bacia do ALTO_JQ. Os preditores são TCI PONDJFM*, VCI PONDJFM*, PDSI PONDJFM*, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_SMA, QSET*_SMA.3.diff. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).....	97
Figura 41-Previsão de vazão Exp4 (QSET*) com 6 meses de antecedência bacia do VACARIA. Os preditores são TCI PONDJFM*, VCI PONDJFM*, PDSI PONDJFM*, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_SMA, QSET*_SMA.3.diff. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).....	98
Figura 42-Previsão de vazão Exp4 (QSET*) com 6 meses de antecedência bacia do Rib_SF. Os preditores são TCI PONDJFM*, VCI PONDJFM*, PDSI PONDJFM*, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_SMA, QSET*_SMA.3.diff. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).....	98
Figura 43-Previsão de vazão Exp4 (QSET*) com 6 meses de antecedência bacia do GONGOGI. Os preditores são TCI PONDJFM*, VCI PONDJFM*, PDSI PONDJFM*, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_SMA, QSET*_SMA.3.diff. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).....	99
Figura 44-Previsão de vazão Exp5(QSET*) com 6 meses de antecedência bacia do ALTO_SF. Os preditores são TCIONDJFM*, VCIONDJFM*, PONDJFM*,	

PDSIONDJFM, PDSI_t-1, PDSIONDJFM_t-2, PDSIONDJFM_t-3, PDSIONDJFM_SMA.3, PDSIONDJFM_SMA.3.diff, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)..... 102

Figura 45-Previsão de vazão Exp5(QSET*) com 6 meses de antecedência bacia do ALTO_JQ. Os preditores são TCIONDJFM*, VCIONDJFM*, PONDJFM*, PDSIONDJFM, PDSI_t-1, PDSIONDJFM_t-2, PDSIONDJFM_t-3, PDSIONDJFM_SMA.3, PDSIONDJFM_SMA.3.diff, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)..... 102

Figura 46-Previsão de vazão Exp5(QSET*) com 6 meses de antecedência bacia do VACARIA. Os preditores são TCIONDJFM*, VCIONDJFM*, PONDJFM*, PDSIONDJFM, PDSI_t-1, PDSIONDJFM_t-2, PDSIONDJFM_t-3, PDSIONDJFM_SMA.3, PDSIONDJFM_SMA.3.diff, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)..... 103

Figura 47-Previsão de vazão Exp5(QSET*) com 6 meses de antecedência bacia do Rib_SF. Os preditores são TCIONDJFM*, VCIONDJFM*, PONDJFM*, PDSIONDJFM, PDSI_t-1, PDSIONDJFM_t-2, PDSIONDJFM_t-3, PDSIONDJFM_SMA.3, PDSIONDJFM_SMA.3.diff, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)..... 103

Figura 48-Previsão de vazão Exp5(QSET*) com 6 meses de antecedência bacia do GONGOGI. Os preditores são TCIONDJFM*, VCIONDJFM*, PONDJFM*, PDSIONDJFM, PDSI_t-1, PDSIONDJFM_t-2, PDSIONDJFM_t-3, PDSIONDJFM_SMA.3, PDSIONDJFM_SMA.3.diff, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)..... 104

Figura 49-Previsão de vazão cruzada Exp6 aplicando-se o modelo SVM obtido para a bacia do ALTO_JQ para gerar previsões na bacia do rio GONGOGI com dados locais e 6 meses de antecedência..... 106

Figura 50-Previsão de vazão cruzada Exp6 aplicando-se o modelo SVM obtido para a bacia do ALTO_JQ para gerar previsões de vazão QSET* na bacia do rio ALTO_SF com dados locais e 6 meses de antecedência..... 106

Figura 51- Previsão de vazão cruzada Exp6 aplicando-se o modelo SVM obtido para a bacia do rio VACARIA para gerar previsões de vazão QSET* na bacia do rio Rib_SF com dados locais e 6 meses de antecedência..... 107

- Figura 52- Previsão de vazão cruzada Exp7 (QSET*) com 6 meses de antecedência na bacia ALTO_SF com modelo e dados da bacia ALTO_JQ. Os preditores são TCIONDJFM, VCIIONDJFM, PDSIONDJFM, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, Q.SET.SMA.3, Q.SET.SMA.3.diff, com uso de modelos ML: máquina de vetor de suporte (SVM linear). As variáveis QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, Q.SET.SMA.3, Q.SET.SMA.3.diff são transferidas da bacia com dados onde foi calibrado o modelo..... 109
- Figura 53- Previsão de vazão cruzada Exp7 (QSET*) com 6 meses de antecedência na bacia GONGOGI com modelo e dados da bacia ALTO_JQ. Os preditores são TCIONDJFM, VCIIONDJFM, PDSIONDJFM, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, Q.SET.SMA.3, Q.SET.SMA.3.diff, com uso de modelos ML: máquina de vetor de suporte (SVM linear). As variáveis QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, Q.SET.SMA.3, Q.SET.SMA.3.diff são transferidas da bacia com dados onde foi calibrado o modelo..... 109
- Figura 54- Previsão de vazão cruzada Exp7 (QSET*) com 6 meses de antecedência na bacia do rio Rib_SF com modelo e dados da bacia VACARIA. Os preditores são TCIONDJFM, VCIIONDJFM, PDSIONDJFM, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, Q.SET.SMA.3, Q.SET.SMA.3.diff, com uso de modelos ML: máquina de vetor de suporte (SVM linear). As variáveis QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, Q.SET.SMA.3, Q.SET.SMA.3.diff são transferidas da bacia com dados onde foi calibrado o modelo..... 110
- Figura 55- Processo de aprendizado desempenhado pelos algoritmos ML..... 116

LISTA DE TABELAS

Tabela 1- Exemplos de aplicações mais comuns de técnicas supervisionadas de Machine Learning por tipo de problema.....	27
Tabela 2-Classificação de intensidade de secas segundo o índice SPI.....	52
Tabela 3-Classificação de intensidade de secas – PDSI.....	54
Tabela 4-Classificação de secas relacionando os índices SPI e VHI.....	56
Tabela 5-Características físicas e climáticas das bacias dos rios selecionados para realização da previsão.....	61
Tabela 6-Variabilidade das vazões (mm·mês-1) no mês de setembro – valores mínimos, máximo, médio, primeiro quartil (1.Q) e terceiro quartil (3.Q) da amostra.....	62
Tabela 7-Estações fluviométricas selecionadas para previsão.....	62
Tabela 8-Disponibilidade de dados no mês de setembro (QSET).....	63
Tabela 9-Descrição dos experimentos realizados em que os modelos são ajustados tendo como entrada as variáveis preditoras e saída a vazão média no mês de setembro (QSET*).....	71
Tabela 10- Resultados experimento Exp1 –fator de inflação da variância (VIF) avaliada pelo modelo de regressão linear (lm) antecipando QSET* com 6 meses de antecedência.....	82
Tabela 11-Resultados experimento Exp2 – Acurácia nas amostras de teste e treino, antecedência de 9 meses na previsão de QSET* com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM), XGBoost (XGB). Preditores utilizados: POND*, QOND*, PDSIOND, VCIOND, TCIOND.....	83
Tabela 12-Resultados experimento Exp2 – Acurácia nas amostras de teste e treino, antecedência de 6 meses na previsão de Q.SET* com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: PONDJFM*, QONDJFM*, PDSIONDJFM, VCIONDJFM, TCIONDJFM.....	86
Tabela 13-Variância explicada pelas funções empíricas ortogonais das variáveis preditoras: precipitação (CHIRPS), TCI, VCI, PDSI.....	90
Tabela 14-Previsão de vazão Exp3 (QSET*) com 6 meses de antecedência bacia do ALTO_SF. Os preditores são as EOF's dos campos de precipitação CHIRPS, TCI, VCI, PDSI. Previsões com uso de modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)	91
Tabela 15-Previsão de vazão Exp4 (QSET*) com 6 meses de antecedência. Os preditores são TCI PONDJFM*, VCI PONDJFM*, PDSI PONDJFM*, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_SMA,	

QSET*_SMA.3.diff. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).....	94
Tabela 16-Importância das variáveis na previsão de 6 meses de antecedência segundo o modelo que apresentou melhor performance SVM-linear no Exp4.....	96
Tabela 17-Previsão de vazão Exp5 (QSET*) com 6 meses de antecedência. Os preditores são TCIONDJFM*, VCIONDJFM*, PONDJFM*, PDSIONDJFM, PDSI_t-1, PDSIONDJFM_t-2, PDSIONDJFM_t-3, PDSIONDJFM_SMA.3, PDSIONDJFM_SMA.3.diff, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).....	100
Tabela 18-Importância das variáveis na previsão de 6 meses de antecedência segundo o modelo que apresentou melhor performance SVM-linear no Exp5.....	101
Tabela 19-Previsão de vazão Exp6 (QSET*) com 6 meses de antecedência. Os preditores são TCIONDJFM, VCIONDJFM, PDSIONDJFM, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, QSET*_t-3, Q.SET.SMA.3, Q.SET.SMA.3.diff, com uso de modelos ML: máquina de vetor de suporte (SVM linear).....	105
Tabela 20- Previsão de vazão cruzada Exp7 (QSET*) com 6 meses de antecedência. Os preditores são TCIONDJFM, VCIONDJFM, PDSIONDJFM, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, Q.SET.SMA.3, Q.SET.SMA.3.diff, com uso de modelos ML: máquina de vetor de suporte (SVM linear). As variáveis QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_t-2, Q.SET.SMA.3, Q.SET.SMA.3.diff são transferidas da bacia com dados onde foi calibrado o modelo.....	108
Tabela A.21-Correlações entre o índice Palmer (PDSI) e o índice de vazão padronizada (SDI).....	132
Tabela A.22- Resultado da análise de variância entre grupos, variável QSET* adimensional.....	133
Tabela A.23-Resultado da análise de variância entre grupos, variável PONDJFM* adimensional.....	133
Tabela A.24-Resultado da análise de variância entre grupos, variável TCI.ONDJFM	133
Tabela A.25-Resultado da análise de variância entre grupos, variável VCI.ONDJFM	134
Tabela A.26-Resultado da análise de variância entre grupos, variável PDSI.ONDJFM	134

LISTA DE ABREVIATURAS E SIGLAS

ALTO_JQ – Área do experimento: bacia do Alto Jequitinhonha no Estado de Minas Gerais

ALTO_SF – Área do experimento: bacia do Alto São Francisco no Estado de Minas Gerais

ANA– Agência Nacional de Águas

ACP– Análise de Componentes Principais

CHIRPS – *Climate Hazards group Infrared Precipitation with Stations*

CPRM – Serviço Geológico do Brasil

CWT – Transformada de Wavelet

EOF – Função empírica ortogonal

GRACE – Gravity Recovery and Climate Experiment

GONGOGI– Área do experimento: bacia do rio Gongogi afluente do rio das Contas no Estado da Bahia

IA- Inteligência Artificial

Lag – intervalo de atraso em uma série temporal

lm – modelo de regressão linear (linear model)

ML- Aprendizado de Máquina (*Machine Learning*)

MK – Teste de Mann-Kendall

NOAA – *National Oceanic Atmospheric Administration*

NDVI – Índice Normalizado da Diferença da Vegetação (*Normalized Difference Vegetation Index*)

p_value – nível de significância de um teste

PDSI- Índice Palmer de Severidade de Seca (*Palmer Drought Severity Index*)

P – Precipitação

POND- Precipitação no trimestre outubro-dezembro

PONDJFM- Precipitação nos trimestres outubro-dezembro e janeiro-março

Pettit – Teste estatístico de Pettit

Qset – Vazão mínima média no mês de setembro

Qset* – Vazão mínima média no mês de setembro adimensionalizada

QOND- Vazão no trimestre outubro-dezembro

QONDJFM- Vazão nos trimestres outubro-dezembro e janeiro-março

RF – Algoritmo Random Forest

Rib_SF– Área do experimento: bacia do ribeirão São Francisco afluente do rio Jequitinhonha no Estado de Minas Gerais

RMSE – Medida de acurácia erro quadrado médio (root mean square error)

R² – grau de correlação associação linear

scPDSI- Índice Palmer de Severidade de Seca auto-calibrado (*Self Calibrated Palmer Drought Severity Index*)

S_{SET} – Armazenamento subterrâneo no final da estação seca

SDI – Índice de anomalia de seca da vazão (*Streamflow Drought Index*)

SNIRH – Sistema Nacional de Informações de Recursos Hídricos

SMA – Componente suavizada de longo prazo do sinal calculada pela média móvel

SMA.diff – Derivada da componente suavizada de longo prazo do sinal calculada pela média móvel

SPI – Índice Padronizado de Precipitação (*Standardized Precipitation Index*)

SVM- Algoritmo Máquina de Vetor de Suporte (*Support Vector Machine*)

SV- Vetor de Suporte (*Support Vector*)

TCI – Índice da Condição da Temperatura (*Temperature Condition Index*)

VACARIA – Área do experimento: bacia do rio Vacaria, afluente do rio Jequitinhonha no estado de Minas Gerais

VCI – Índice da Condição da Vegetação (*Vegetation Condition Index*)

WCT- Transformada Wavelet-Coherence

XGBoost – Algoritmo *Extreme Gradient Boosting*

XGB – Abreviação de XGBoost

SUMÁRIO

1	INTRODUÇÃO	22
1.1	OBJETIVOS.....	25
1.1.1	OBJETIVO GERAL.....	25
1.1.2	OBJETIVOS ESPECÍFICOS	25
2	REFERENCIAL TEÓRICO.....	27
2.1	APRENDIZADO ESTATÍSTICO (MACHINE LEARNING)	27
2.1.1	O QUE É MACHINE LEARNING?.....	27
2.1.2	APRENDIZAGEM SUPERVISIONADA E NÃO-SUPERVISIONADA	30
2.2	ERROS ENVOLVENDO MODELOS	31
2.3	GENERALIZAÇÃO.....	33
2.4	ALGORITMOS MACHINE LEARNING.....	34
2.4.1	MÁQUINAS DE VETOR DE SUPORTE (SVM)	35
2.4.2	RANDOM FOREST (RF)	41
2.4.3	BOOSTING XGBOOST.....	43
2.5	TRANSFORMAÇÃO DE DADOS	47
2.6	INDICES DE SECA	50
2.6.1	ÍNDICE DE PRECIPITAÇÃO PADRONIZADA (SPI).....	50
2.6.2	ÍNDICE DE SEVERIDADE DE SECA PALMER (PDSI)	52
2.6.3	NOAA - AVHRR VEGETATION CONDITION INDEX (VHI).....	54
3	MATERIAIS E MÉTODOS.....	57
3.1	DESCRIÇÃO GERAL.....	57
3.2	ÁREA DE ESTUDO	58
3.3	DADOS	62
3.3.1	DADOS HIDROLÓGICOS.....	62
3.3.2	DADOS DE PRECIPITAÇÃO.....	63
3.3.3	ÍNDICE PALMER.....	63
3.3.4	ÍNDICES VCI E TCI.....	64
3.4	ANÁLISE DE TENDÊNCIAS.....	65
3.5	ANÁLISE DE CORRELAÇÕES	66
3.6	MODELAGEM ML	67
3.6.1	TRANSFORMAÇÕES DE DADOS	67
3.6.2	ALGORITMOS EMPREGADOS.....	67
3.6.3	DESCRIÇÃO DOS EXPERIMENTOS	68
3.6.4	SEPARAÇÃO DA AMOSTRA DE TREINO E VALIDAÇÃO	72
3.6.5	CALIBRAÇÃO.....	73

4	RESULTADOS	75
4.1.1	ANÁLISE DE ESTACIONARIEDADE E DE MUDANÇA DE TENDÊNCIAS.....	75
4.1.2	ANÁLISE DE CORRELAÇÃO	77
4.1.3	RESULTADOS DOS EXPERIMENTOS	82
5	DISCUSSÃO	111
5.1	CARACTERÍSTICAS DO SINAL DA VAZÃO MÍNIMA Q_{SET}	111
5.2	OS PREDITORES VCI, TCI.....	112
5.3	PALMER (PDSI).....	113
5.4	A VARIABILIDADE ESPACIAL	114
5.5	O QUE OS ALGORITMOS APREENDERAM SOBRE HIDROLOGIA?	114
5.6	PERFORMANCE E GENERALIZAÇÃO.....	116
5.7	O PAPEL DAS CIÊNCIAS HIDROLOGICAS NA ERA DO MACHINE LEARNING	118
6	CONCLUSÕES E RECOMENDAÇÕES	119
6.1	CONCLUSÕES	119
6.2	RECOMENDAÇÕES.....	120
	REFERÊNCIAS	121
	APÊNDICE A– TESTES ESTATÍSTICOS	132

1 INTRODUÇÃO

A premissa de estacionariedade hidrológica adotada nas últimas décadas se tornou inadequada para o exercício da engenharia e gerenciamento dos recursos hídricos, considerando que os padrões de mudança são complexos, as incertezas são largas e a base de conhecimento tem se modificado rapidamente (MILLY et al., 2008). No gerenciamento de riscos de seca, desconsiderar esses fatores leva a avaliações enviesadas por desconsiderar as mudanças aceleradas do clima, já significativas nos tempos atuais (HOYLMAN; BOCINSKY; JENCISO, 2022).

No Brasil, no período de 2011-2019, eventos de seca foram observados em todo o território, sendo mais severos e abrangentes no período de 2011-2017. O ano hidrológico 2014-2015 se destacou pela elevada ocorrência de eventos de seca moderados a severos em todo país (CUNHA et al., 2019). Durante o período de 2011-2016 a região nordeste do Brasil registrou uma seca com intensidade e abrangência ainda sem precedentes (BRITO et al., 2018; MARENGO ORSINI et al., 2018).

Projeções futuras apontam mudanças nos padrões sazonais de precipitação com decréscimo de precipitação no norte da bacia do Rio da Prata (GOMES et al., 2022). Nos próximos 30 anos, a região sudoeste do Brasil deve enfrentar condições sem precedentes de seca hidrológica independente dos níveis de emissão globais (SATO et al., 2022). No bioma Cerrado, a maior savana tropical do mundo, projeções apontam para ocorrência de secas meteorológicas e hidrológicas mais intensas, frequentes, de maior duração e mais severas, com secas hidrológicas mais longas que as meteorológicas e com reduções de vazão (RODRIGUES et al., 2020).

Além dos efeitos climáticos, acrescentam-se as intervenções antrópicas com evidentes impactos nas depleções de reservas subterrâneas (GONÇALVES et al., 2020) e superficiais (DA SILVA et al., 2021) estratégicas para o país, com grandes consequências sobre a disponibilidade hídrica em período de estiagem. Entre as quais o aumento do risco de ocorrências de vazões mínimas extremas, aumento da frequência e redução das vazões mínimas em termos médios (OUYANG, 2012).

Ainda diante do fato de que 84% da população brasileira se abastece a partir de mananciais superficiais, a gestão da disponibilidade hídrica no Brasil ainda se encontra fortemente alicerçada sobre estimativa de vazões mínimas de referência (ANA, 2020), normalmente calculadas a partir de séries históricas ou por estimativas regionais que não

consideram, além da não-estacionariedade, a sazonalidade ou padrões hidrológicos locais (SERRANO et al., 2020).

No cenário de não-estacionariedade induzida pelo clima ou pressões antrópicas, as previsões de longo prazo de vazões mínimas são relevantes, uma vez que permitem: antecipar a seca hidrológica e induzir sem atraso a mudança no padrão de consumo de água (SOUSA ESTÁCIO et al., 2022); contribuir com a mitigação do custo da escassez hídrica através da realocação de água entre usos e redução dos déficits hídricos, possibilitando a elaboração de estratégias ótimas com maior antecedência (MATTIUZI; MARQUES; MEDELLÍN-AZUARA, 2019); antecipar desdobramentos tais como secas socioeconômicas e a ocorrência simultânea de outros desastres relacionados à seca hidrológica (GETIRANA; LIBONATI; CATALDI, 2021), além de contribuir com a ampliação do instrumento de alocação de água negociada em sistemas hídricos.

Previsões hidrológicas com mais de um mês de antecedência são consideradas de longo prazo e são realizadas principalmente de duas formas: (1) acoplamento de modelos de simulação hidrológica a modelos de circulação global (WOOD et al., 2002; TUCCI et al., 2003; MAHANAMA et al., 2008; SHUKLA; LETTENMAIER, 2011; MAHANAMA et al., 2012; VAN DIJK et al., 2013; PETRY et al., 2021); e (2) modelos estatísticos treinados sobre índices climáticos, meteorológicos e variáveis hidrológicas (MAURER; LETTENMAIER, 2003; SEIBERT; MERZ; APEL, 2017; DE ARAÚJO et al., 2020; DE PAIVA; MONTENEGRO; CATALDI, 2020).

Modelos distribuídos de simulação hidrológica requerem significativamente mais dados de entrada e maior esforço computacional (WOOD; ROUNDY; et al., 2011; SEIBERT; MERZ; APEL, 2017). Por outro lado, os modelos estatísticos requerem menos dados e não necessitam de medidas físicas, fundamentando-se em algoritmos de aprendizado de máquina (*machine learning*). Nesse contexto, as redes neurais artificiais-ANN (OJHA; ABRAHAM; SNÁŠEL, 2017) e as máquinas de vetor de suporte - SVM (RAGHAVENDRA; DEKA, 2014) são os algoritmos de aprendizado mais utilizados desde o início dos anos 90 no campo das ciências hidrológicas.

Por décadas, as aplicações de ANN nas previsões hidrológicas não passaram de estudos repetitivos com o objetivo de testar diferentes conjuntos de dados, métodos de otimização, e refinamento de estruturas, sem grandes progressos científicos, o que levou tal técnica a ser vista com ceticismo pelos hidrólogos, devido à falta de significado físico dos seus resultados (ABRAHART et al., 2012) e contribuindo para que as ferramentas

de inteligência artificial (AI) na hidrologia fossem conhecidas como “caixas pretas” (SAHOO et al., 2019) cujos modelos apresentam baixo poder de generalização.

Nos anos recentes tipos especiais de arquiteturas de ANN ganharam destaque com o nome de aprendizado profundo (*Deep Learning*) no campo das ciências hidrológicas por apresentar características promissoras, tais como: a capacidade de aprendizado sobre dependências temporais de longo termo; habilidade em previsões de longo prazo; capacidade de encontrar sinergia entre diferentes bases de dados da mesma variável hidrológica; grande poder de aprendizado em variações espaço-temporais; grande capacidade de generalização; e performance superior a modelos hidrológicos e a modelos *machine learning* convencionais (KRATZERT et al., 2018; SAHOO et al., 2019; NEARING et al., 2021; KRATZERT et al., 2021).

Entretanto, o *deep learning* ainda encontra resistência para ser absorvido no campo das ciências hidrológicas (NEARING et al., 2021). Embora caminhos para promover maior integração e aceitação pelos profissionais e cientistas de hidrologia tenham sido apontados (SHEN et al., 2018; KARPATNE et al., 2017), o salto dado pelo *deep learning* foi bastante disruptivo, principalmente para uma comunidade técnica e científica que não acompanhou sua evolução. Assim, a previsão de fenômenos hidrológicos com emprego de algoritmos de alta performance como o *Extreme Gradiante Boosting* - XGBoost (CHEN; GUESTRIN, 2016) e de grande poder de generalização como o clássico SVM irá provavelmente contribuir para uma abertura gradual e aceitação do *machine learning*.

A Hidrologia ainda está assimilando a premissa de não-estacionariedade (MILLY et al., 2008) e o fato de quase três décadas frustradas de tentativas de operacionalização de previsões com base em *machine learning* (ABRAHART et al., 2012). Desta forma, a comunidade técnica e científica do campo das ciências hidrológicas precisa ser convencida e se apropriar das bases técnicas nestes tipos de previsões para assim se sentir segura na sua aplicação.

Portanto, ainda há bastante espaço para a aplicação de algoritmos clássicos de *machine learning* buscando-se generalização em problemas relevantes e de natureza operacional na hidrologia. Um exemplo é a previsão de longo prazo de vazões mínimas em períodos de estiagem, em situações de não-estacionariedade, baseados em conhecimento físico teórico extraído a partir de sua aplicação em massas de dados produzidas principalmente pelas técnicas de sensoriamento remoto - SR (MAXWELL; WARNER; FANG, 2018; AGHAKOUCHAK et al., 2015).

Até o presente momento, não se tem conhecimento de nenhuma aplicação de previsão envolvendo modelos *machine learning* que tenham empregado os índices de seca *Temperature Condition Index*-TCI, *Vegetation Condition Index*-VCI, *Vegetation Health Index*-VHI (KOGAN, 1995; KOGAN, 1997; KOGAN, 2002; YANG; KOGAN; GUO, 2020) obtidos por meio de SR pelos sensores AVHRR e VIIRS da NOAA, para previsão de vazões mínimas. E raros são os trabalhos como o de KIM&VALDÉS (2003), que utilizaram o índice meteorológico de Palmer (PDSI) nesse tipo de previsão.

Este trabalho vem preencher essa lacuna, e teve como objetivo a realização de previsões de longo prazo de vazões mínimas com 6 e 9 meses de antecedência, em cinco bacias hidrográficas localizadas no Cerrado e na Mata-Atlântica, nas quais investigações prévias mostraram situações de não-estacionariedade climática e hidrológica, especialmente devido à recente seca severa, fazendo uso de modelos estatísticos derivados por algoritmos *machine learning* e treinados sobre dados disponibilizados em escala global e produtos derivados de sensoriamento remoto.

1.1 OBJETIVOS

1.1.1 OBJETIVO GERAL

Avaliar a habilidade de algoritmos ML na previsão de vazões mínimas de longo prazo, em situações de não-estacionariedade, utilizando como preditores: índices de seca e produtos derivados de sensoriamento remoto, disponibilizados em escala global com média e baixa resolução.

1.1.2 OBJETIVOS ESPECÍFICOS

- Investigar a habilidade de aprendizagem dos algoritmos de ML sobre processos hidrológicos relacionados às vazões mínimas a partir de observações indiretas de variáveis representativas do ciclo hidrológico a partir de sensoriamento remoto;
- Avaliar a habilidade dos índices de seca VCI (*Vegetation Condition Index*) e TCI (*Temperature Condition Index*) disponibilizados pela AVHRR/VIIRS-NOAA na previsão de vazões mínimas e suas respectivas correlações com variáveis hidrológicas;
- Investigar a habilidade de previsão do índice Palmer (PDSI) obtido em escala global, como substituto da vazão em bacias sem informações;
- Avaliar o poder de generalização de modelos gerados por algoritmos de ML;
- Avaliar o desempenho de algoritmos de ML em situações de mudança de tendência;

- Avaliar a capacidade da ML treinada sobre produtos globais em realizar previsão de vazões antes e durante desastres extremos como o evento de seca de 2010-2016;

2 REFERENCIAL TEÓRICO

2.1 APRENDIZADO ESTATÍSTICO (MACHINE LEARNING)

2.1.1 O que é Machine Learning?

Aprendizado de Máquina (*machine learning*) é o processo de aprendizado dos algoritmos a partir dos dados, que utiliza o conhecimento adquirido através da experimentação para aprimorar sua performance e posteriormente subsidiar futuras decisões. Consiste em uma técnica extremamente poderosa que se tornou rapidamente o mecanismo propulsor da nova economia baseada em dados (BRINK et al., 2017) com diversas aplicações práticas, como resumidamente se pode exemplificar na **Tabela 1**.

Em geral, os diversos algoritmos de ML são utilizados de forma a gerar classificadores para um conjunto de exemplos, atribuindo-se uma determinada informação, o rótulo da classe à qual ela pertence, empregando um processo de indução de um classificador a partir de uma amostra de treinamento, que deve ser capaz de prever classes de instâncias quaisquer, do domínio em que ele foi treinado (LORENA; DE CARVALHO, 2007).

Tabela 1- Exemplos de aplicações mais comuns de técnicas supervisionadas de *Machine Learning* por tipo de problema

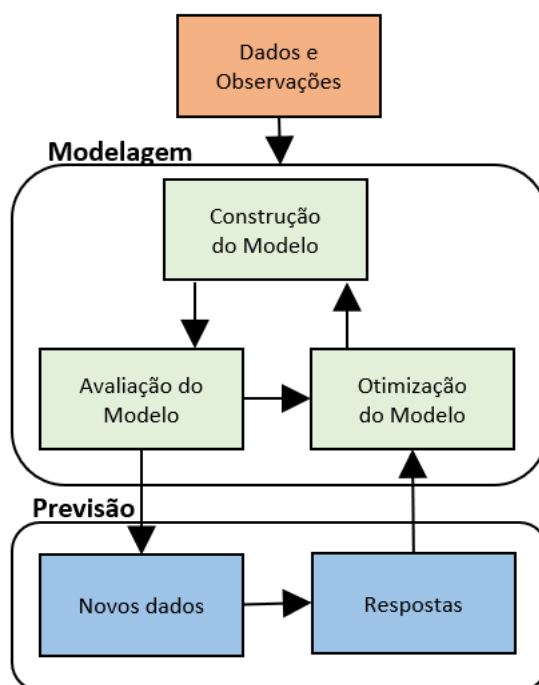
Problema	Descrição	Exemplos de aplicações
Classificação	Determinar a classe discreta à qual um indivíduo pertence, baseando-se em suas características como dados de entrada.	Filtros de SPAM, Análise de sentimento, detecção de fraudes, previsão de rotatividade, suporte em casos de sinalização, personalização de conteúdo, detecção de defeitos em processos manufaturados, segmentação de clientes, descoberta de eventos, genoma, eficácia de medicamentos
Regressão	Prever um valor (número real) baseando-se em dados de entrada	Previsão de estoques em mercado, previsão de demandas, estimativa de preços, otimização de pedidos e encomendas, previsão do tempo
Recomendação	Prever quais alternativas um usuário está disposto a escolher	Recomendação de produtos, processos de recrutamento de pessoal, recomendações de séries e filmes em plataformas de TV, sites de relacionamento, recomendações de conteúdo
Imputação	Inferir sobre valores faltantes a partir de dados de entrada	Casos de ausência de informação em registros médicos e cadastros comerciais, dados de censo.

Fonte: adaptado de BRINK et al., (2017).

Segundo PROVOST & FAWCETT (2016) o ML surgiu como um subcampo da Inteligência Artificial, que estava preocupada com métodos para melhorar o conhecimento ou desempenho de um agente inteligente ao longo do tempo e a resposta à experiência do agente no mundo. O Campo de Mineração de Dados (*Data Mining*) se iniciou como uma ramificação do ML, entretanto mais orientada à preparação de dados, modelos de aprendizagem e avaliação, não se preocupando da forma como o agente utilizará o conhecimento adquirido para agir em seu ambiente – que são preocupações do ML. Entretanto, técnicas e algoritmos são compartilhados entre os dois campos.

De forma geral a construção de um modelo ML segue o fluxograma de etapas apresentado na **Figura 1** que consiste nas etapas de preparação dos dados, construção, avaliação e treinamento do modelo e aplicação em novos dados.

Figura 1-Fluxo de um modelo típico de *Machine Learning*. A partir de dados e observações históricas se constrói um modelo utilizando um algoritmo ML. Em seguida se faz a avaliação de performance e se otimiza a acurácia e escalabilidade para se ajustar aos requisitos da modelagem. Com o modelo ajustado se realiza a previsão sobre novos dados.

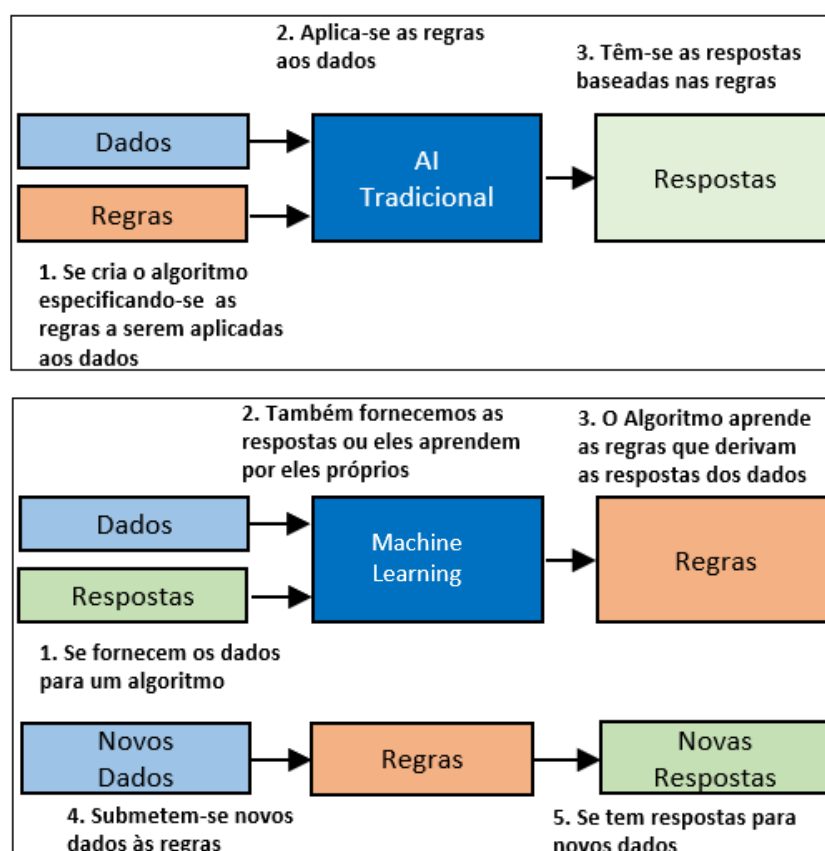


Fonte: adaptado de BRINK et al., (2017).

O Aprendizado de Máquina (*Machine learning*) também referenciado muitas vezes como aprendizado estatístico (*Statistical Learning*) é um subcampo da Inteligência Artificial (AI) onde os algoritmos “aprendem” padrões extraídos a partir dos dados para desempenhar tarefas específicas. O termo “aprendizado” merece destaque, pois é o que diferencia o *Machine Learning* das abordagens tradicionais da AI (RHYS, 2020).

A AI tradicional é programática, ou seja, se fornece ao computador uma série de regras para que ele encontre os novos dados. O problema dessa abordagem é que ela prescinde do conhecimento prévio de todas as saídas (resultados) possíveis que o programa pode gerar. Em contraste, na abordagem ML ao invés de se fornecer as regras, são fornecidos os dados para que ele derive as regras por si mesmo. A vantagem esta abordagem é que a máquina pode “aprender” padrões que não se sabia sequer existir nos dados. E quanto mais dados são fornecidos, melhor seu aprendizado e melhores padrões são extraídos (RHYS, 2020). A diferença entre essas abordagens é exemplificada na **figura 2.2**.

Figura 2-AI tradicional vs. *Machine Learning* AI. Nas aplicações de AI tradicional se fornece ao computador um conjunto de regras. Quando se fornecem os dados, ele fornece as respostas relevantes. Na abordagem Machine Learning se fornecem os dados ao computador e as respostas, então o computador aprende as regras por ele mesmo. Quando se fornecem novos dados ele utiliza essas regras para fornecer as novas respostas.



Fonte: Modificado de (RHYS, 2020)

MITCHELL (1997) define o ML como qualquer programa computacional que aprimore sua performance em alguma tarefa através da experiência, mais precisamente fornece a seguinte definição: Um programa de computador é o algoritmo ML que recebe instruções para aprender com a experiência E no que diz respeito a

algumas classes de tarefa **T** e medidas de performance **P**, se essa performance nas tarefas **N**, conforme medidas por **P**, se aperfeiçoam através da experiência **E**.

2.1.2 *Aprendizagem Supervisionada e Não-Supervisionada*

Técnicas de ML podem ser classificadas principalmente em duas categorias: supervisionadas e não-supervisionadas (KASSAMBARA, 2019).

As supervisionadas consistem em construir modelos matemáticos para previsão de futuras observações, e são classificadas em dois grupos principais:

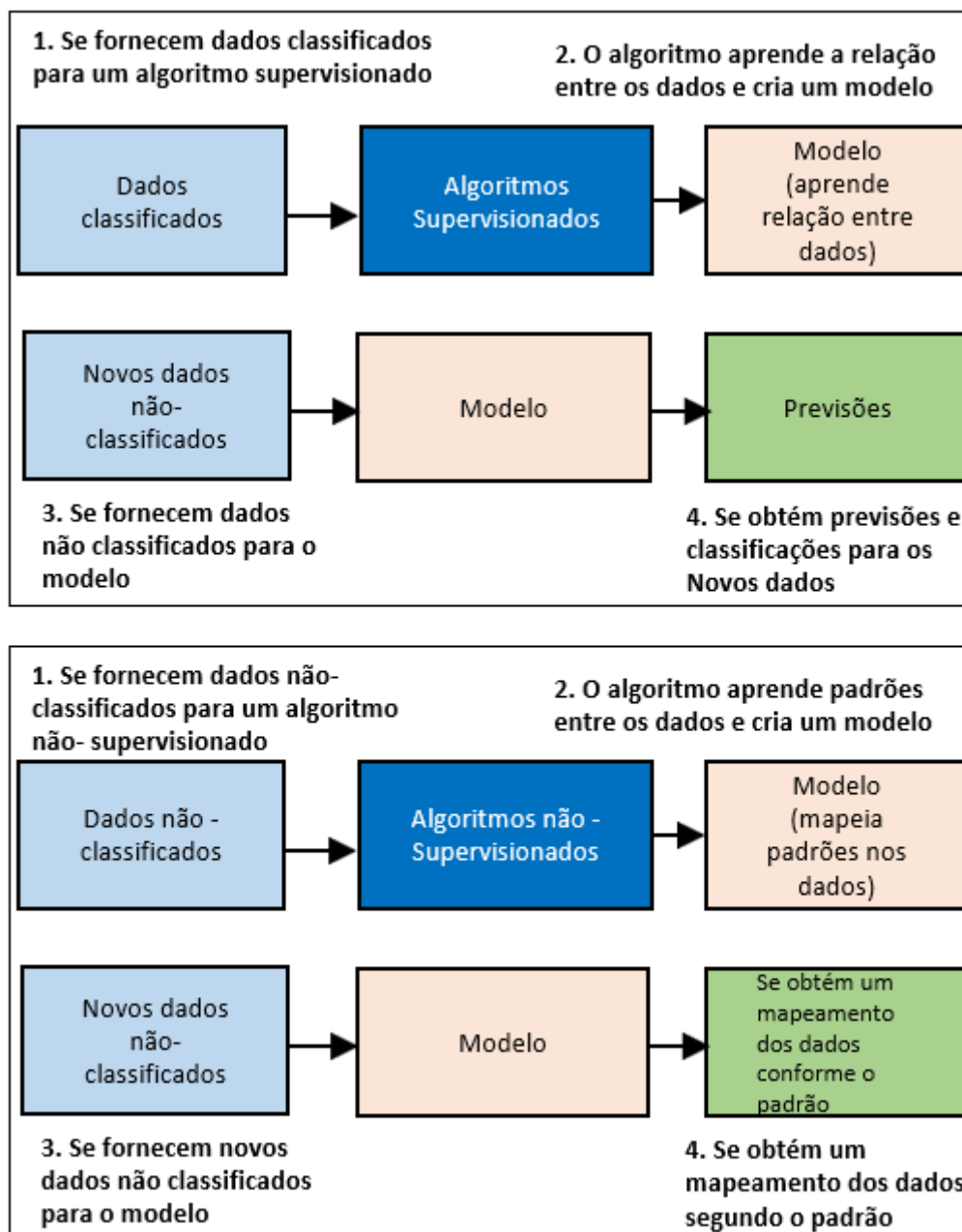
- **Regressão**, normalmente utilizados para previsão de variáveis contínuas;
- **Classificação**, usados para prever classes (ou grupos) ao qual os indivíduos pertencem. Também se incluem nesse grupo os modelos que preveem a probabilidade de indivíduo pertencer a uma classe.

Esses métodos são supervisionados porque o modelo é construído a partir de respostas conhecidas, isto é, ML aprende a partir das observações para prever casos futuros.

Os métodos não-supervisionados têm o objetivo de identificar padrões ou grupos com grande similaridade entre os membros. Neste caso os algoritmos aprendem como agrupar ou obter as informações mais importantes do conjunto de dados.

A **Figura 3** ilustra a diferença entre os dois métodos. No método supervisionado os algoritmos tomam como verdade os dados já classificados e constroem um modelo preditivo para prever classes em um novo conjunto de dados ainda não-classificados. Os algoritmos não-supervisionados trabalham a partir dados não classificados e extraem seus padrões, de forma que um novo conjunto de dados pode ser mapeado a partir deles (RHYS, 2020).

Figura 3-Funcionamento dos algoritmos supervisionados versus não-supervisionados. Os supervisionados constroem um modelo a partir de dados classificados. Os não-supervisionados extraem padrões a partir de dados não-classificados.



Fonte: Modificado de (RHYS, 2020)

2.2 ERROS ENVOLVENDO MODELOS

A construção de um modelo consiste em descobrir a relação entre as respostas Y e os p preditores $X_1, X_2, X_3, \dots, X_p$, onde é possível assumir que a relação entre Y e $X = (X_1, X_2, X_3, \dots, X_p)$, pode ser escrita em sua forma mais geral como

$$Y = f(X) + \varepsilon$$

Equação 1

Onde f é uma função desconhecida sobre $X_1, X_2, X_3, \dots, X_p$ e ε é um termo aleatório que se refere ao Erro, independentemente do valor de X e que possui média zero. Nesta formulação, f representa a informação sistemática de X sobre Y . Em essência, o aprendizado estatístico se refere ao conjunto de abordagens e métodos para estimativa de f , que se quer determinar normalmente por dois motivos: previsão ou inferência (JAMES et al., 2013).

Associados à modelagem existem dois tipos de erros: os que são reduzíveis e os não-reduzíveis. Os reduzíveis podem ser minimizados através da aplicação de uma técnica apropriada no processo de modelagem, entretanto mesmo uma perfeita estimativa de f estaria sujeita a erros uma vez que não se pode reduzir a parcela ε .

Considerando uma dada estimativa de $\hat{f}$ e um conjunto de preditores X , os quais produzem as previsões $\hat{Y} = \hat{f}(X)$, e fixando-se $\hat{f}$ e X os erros podem ser postos conforme Equação 2. Onde $E[f(X) - \hat{f}(X)]^2$ é a parcela redutível do erro e $Var(\varepsilon)$ a parcela não redutível que se trata da variância associada ao erro.

$$E[(Y - \hat{Y})]^2 = E[f(X) + \varepsilon - \hat{f}(X)]^2 = E[f(X) - \hat{f}(X)]^2 + Var(\varepsilon) \quad \text{Equação 2}$$

A construção de um modelo visa minimizar a parcela redutível do erro e a parcela não redutível coloca a variação do erro dentro de um patamar de incerteza.

A partir da **Equação 2** pode-se derivar uma expressão para o erro esperado de um modelo $\hat{Y} = \hat{f}(X)$ em um ponto $X = x_0$ pela expressão

$$\text{Err}(x_0) = E \left[(Y - \hat{f}(x_0))^2 \mid X = x_0 \right] = \sigma_\varepsilon^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) =$$

Equação 3

$$\text{Erro Irredutível} + \text{Bias}^2 + \text{Variância}$$

Equação 4

O primeiro termo é a variância do alvo em torno da sua média verdadeira $f(x_0)$ e não pode ser removida não importa a quão bom seja o modelo empregado. O segundo termo é o viés (Bias) elevado ao quadrado, a quantidade média dos desvios da

estimativa em relação à média verdadeira, e o último termo é a variância, os quadrados dos desvios das estimativas $\hat{f}(x_0)$ em torno da sua média. Quanto mais complexo for o modelo $\hat{f}$, menor será o viés e maior será a variância (HASTIE; TIBSHIRANI; FRIEDMAN, 2017).

O erro devido ao viés ocorre pelas simplificações assumidas pelo modelo em representar o fenômeno. Geralmente em aplicações de modelos paramétricos se têm um viés mais elevado em troca de um aprendizado mais ágil e mais simples, porém menos flexível. Por outro lado, se obtém uma performance preditiva menor em problemas complexos (BROWNLEE, 2019).

A variância é a quantidade de erro decorrente da utilização do modelo em dados distintos da amostra de treino. Considerando que o modelo é estimado a partir da amostra de treino, é de se esperar que o algoritmo apresente alguma variância. Algoritmos que naturalmente apresentam uma variância elevada são fortemente influenciados pelos dados utilizados na fase de treinamento (BROWNLEE, 2019).

2.3 GENERALIZAÇÃO

No campo do *machine learning* o aprendizado da função através do treino sobre os dados é também chamado como aprendizado por indução, o que significa que diferentemente dos métodos dedutivos, conceitualmente o aprendizado acontece do específico para o geral. Generalização se trata de quão bem os conceitos e habilidades aprendidas por um modelo ML se aplicam a dados não utilizados ou não vistos pelo modelo durante a fase de treinamento. De forma geral um modelo ML é considerado bom quando é capaz de generalizar a partir da sua amostra de treino para quaisquer dados do domínio do problema (BROWNLEE, 2019).

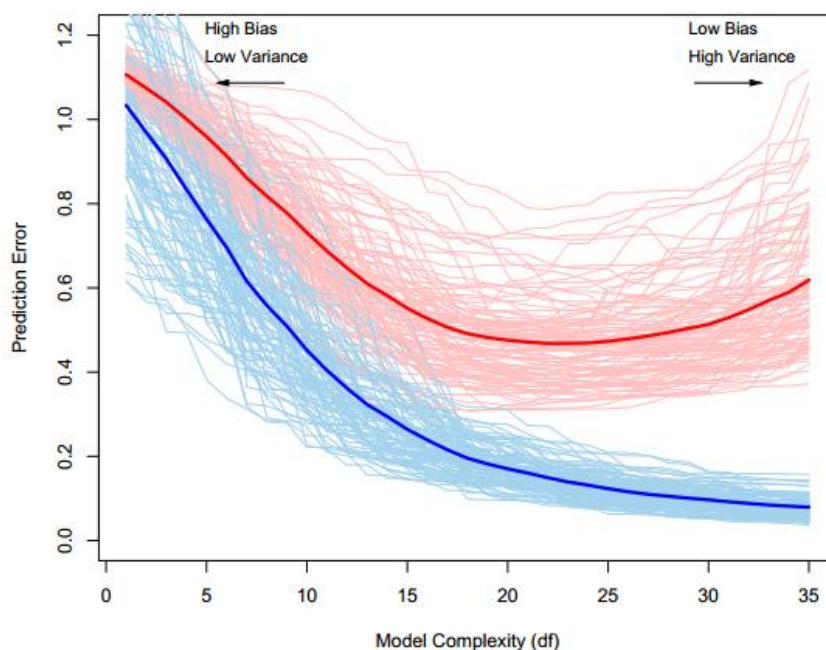
A capacidade de generalização de um modelo diz respeito a sua habilidade de previsão em dados de teste não utilizados no treinamento. A avaliação dessa métrica de performance é extremamente importante em termos práticos, uma vez que se trata do aspecto final de maior peso na escolha do modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2017).

Na **Figura 4** é ilustrado os resultados de um experimento que exemplificam a variação da capacidade de generalização em função da complexidade e a negociação entre o viés e variância que ocorre durante a modelagem.

Segundo BROWNLEE (2019), trata-se de uma escolha que não se pode escapar no campo do ML: aumentar o viés e reduzir a variância ou diminuir o viés e

aumentar a variância. Essa negociação está fortemente relacionada aos algoritmos utilizados.

Figura 4- Análise comparativa entre o erro em uma amostra de treino e a amostra de teste variando-se o grau de complexidade do modelo. As linhas azuis se referem aos erros na amostra de treino e as linhas vermelhas se referem ao erro na amostra de teste. Experimento realizado em 100 amostras de treino de tamanho 50 cada. As linhas sólidas ilustram os erros esperados.



Fonte: HASTIE; TIBSHIRANI; FRIEDMAN (2017).

2.4 ALGORITMOS MACHINE LEARNING

Um algoritmo *machine learning* é um conjunto de regras que a máquina segue ou emprega para aprender como atingir um objetivo particular. A saída deste algoritmo é um modelo *machine learning*. Tais algoritmos também são chamados de indutores (MOLNAR, 2018).

Dentre os diversos tipos de algoritmos existentes, aqui são abordados três algoritmos ML escolhidos para serem empregados neste experimento: Máquinas de Vetor de Suporte (ou *Support Vector Machine* – SVM); *Random Forest* (RF) e *Extreme Gradient Boost* - XGBoost.

2.4.1 Máquinas de Vetor de Suporte (SVM)

SVM é um dos algoritmos de ML mais empregados no campo das ciências hidrológicas, com aplicações em previsões de cheia (MOSAVI; OZTURK; CHAU, 2018), previsão da umidade do solo (LEE; KIM; LEE, 2013), previsão de descarga de sedimentos; geração de curva chave (SIVAPRAGASAM; MUTTIL, 2005), simulação chuva-vazão (DIBIKE et al., 2001), previsão de longo prazo de níveis de reservatórios (KHAN; COULIBALY, 2006), previsão de longo-prazo de vazões (LIN; CHENG; CHAU, 2006), dentre outras (RAGHAVENDRA; DEKA, 2014).

SVM foi desenvolvida sobre a teoria do aprendizado estatístico por CORTES; VAPNIK; SAITTA, 1995; VAPNIK (1995) e fundamenta-se na hipótese de minimização do risco estrutural e não somente do risco empírico, no sentido a buscar uma maior capacidade de generalização, uma vez que uma boa acurácia na amostra de treino não necessariamente garante boa acurácia na fase de testes.

A minimização do risco estrutural segue a teoria da convergência em probabilidade uniforme desenvolvida em 1974 por Vapnik e Chervonenkis (VC) que estabelece limites para os desvios de risco empírico (Equação 5) e do risco esperado (Equação 22). Essa teoria mostra que é crucial se restringir a classe de funções que o algoritmo pode implementar em função do tamanho da amostra de treino. O parâmetro h na Equação 22 é a dimensão VC de um conjunto de funções e descreve a capacidade destas funções em representar o conjunto de dados n , com a confiabilidade $(1 - \theta)$. A dimensão VC é uma medida da complexidade do modelo e é normalmente proporcional ao número de parâmetros do classificador f . Particularmente quando a relação $\frac{n}{h}$ é pequena, um pequeno risco empírico (Equação 5) não garante a um pequeno valor no risco real e nesse caso deve-se controlar o risco pela dimensão do VC do modelo no lado direito da Equação 22 (DIBIKE et al., 2001).

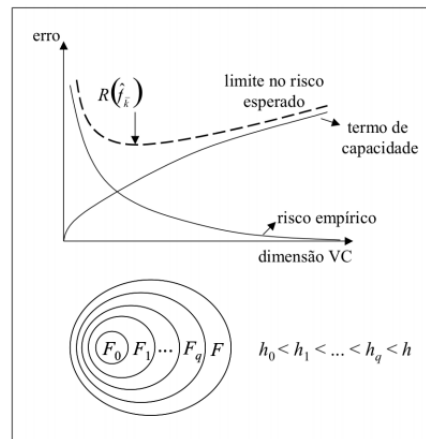
A contribuição desta inequação (Equação 22) consiste em mostrar ser possível controlar a capacidade de um conjunto de funções F do qual o classificador f é extraído, e ainda é possível minimizar esse risco pela escolha adequada (LORENA; DE CARVALHO, 2007).

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} \times |y - f(x)| \quad \text{Equação 5}$$

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h \times (\ln(\frac{2 \cdot n}{h} + 1) - \ln(\frac{\theta}{4}))}{n}} \quad \text{Equação 6}$$

Os conceitos discutidos são apresentados na **Figura 12**, que ilustra o princípio de minimização do risco estrutural.

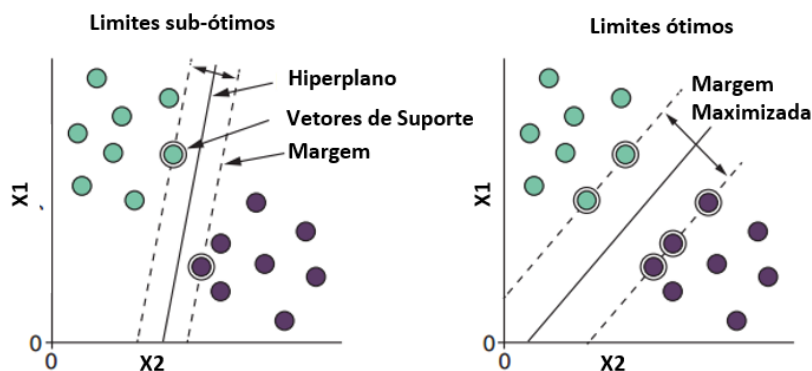
Figura 5. Princípio de minimização do risco estrutural



Fonte: LORENA; DE CARVALHO (2007)

O objetivo principal do SVM é estabelecer uma fronteira de decisão em um espaço de características que separe os dados em diferentes classes, através do posicionamento ótimo de um hiperplano entre as classes que minimiza o erro de generalização e maximiza a margem. Os pontos tocados pela margem são denominados vetores de suporte, que são os pontos de treinamento mais importantes, uma vez que definem o hiperplano e a linha decisória (**Figura 6**).

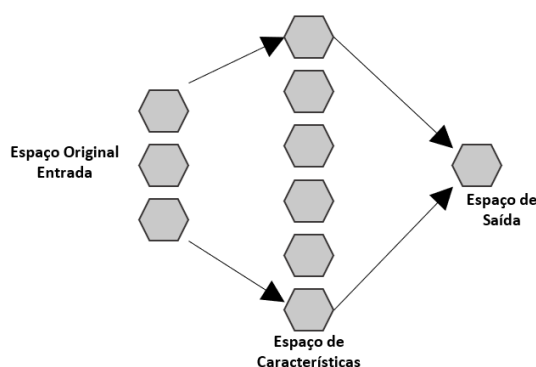
Figura 6- Ilustração do processo de otimização em que o algoritmo SVM procura um hiperplano (linha sólida) que atravessa o espaço de características. O hiperplano ótimo é o que maximiza a margem (linhas tracejadas). A margem é a região no entorno do hiperplano que toca o mínimo de elementos possível, esses elementos que definem a margem são chamados de vetores de suporte ilustrados com contorno nos círculos.



Fonte: RHYS (2020) modificado pelo autor

Nos casos em que os dados não são linearmente separáveis, SVM lança mão de um recurso que consiste num mapeamento dos dados originais num espaço de maiores dimensões, denominado de espaço de características ou espaço de Hilbert, conforme **Figura 7**. com o emprego de funções não-lineares também chamadas de funções kernel.

Figura 7- O algoritmo SVM mapeia os dados de entrada vindo do espaço original num espaço de maiores dimensões chamado espaço de características

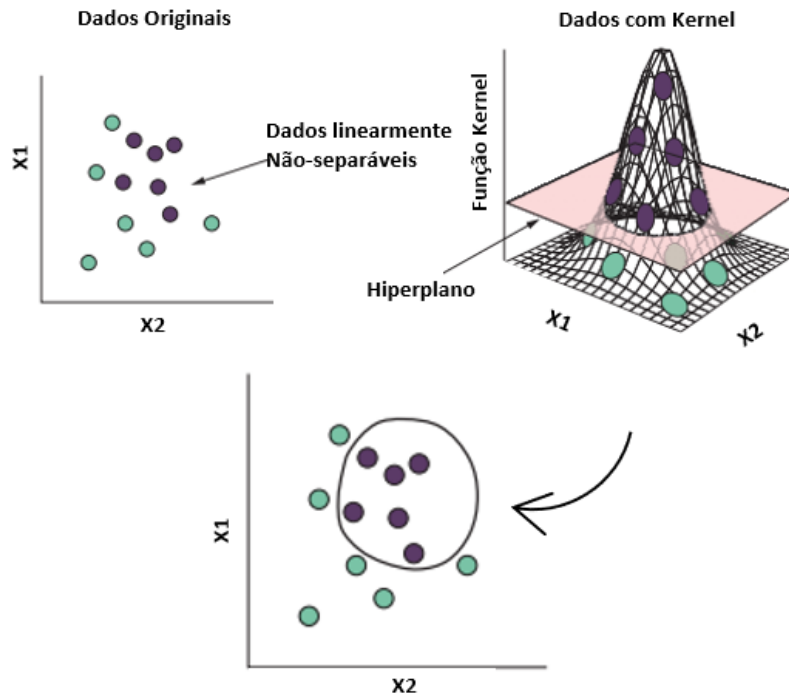


Fonte: RAGHAVENDRA; DEKA (2014) modificado pelo autor

Os kernels apresentam a vantagem de operar nos dados originais e espaço de entrada, onde o resultado da classificação é a soma ponderada das funções kernel calculadas nos vetores de suporte (RAGHAVENDRA; DEKA, 2014). Esse processo

também pode ser compreendido como uma deforma ou distorção do espaço de características, sendo que o Kernel consiste na dimensão adicional (RHYS, 2020) conforme ilustrado na **Figura 8**.

Figura 8- Ilustração da aplicação de um Kernel onde o algoritmo SVM adiciona uma dimensão a mais para realizar a separação dos dados não linearmente separáveis em seu espaço original. A dimensão adicional no espaço de características permite com que os dados sejam separados por um hiperplano. Quando os dados são projetados novamente em duas dimensões o hiperplano aparece como um limite de decisão curva.



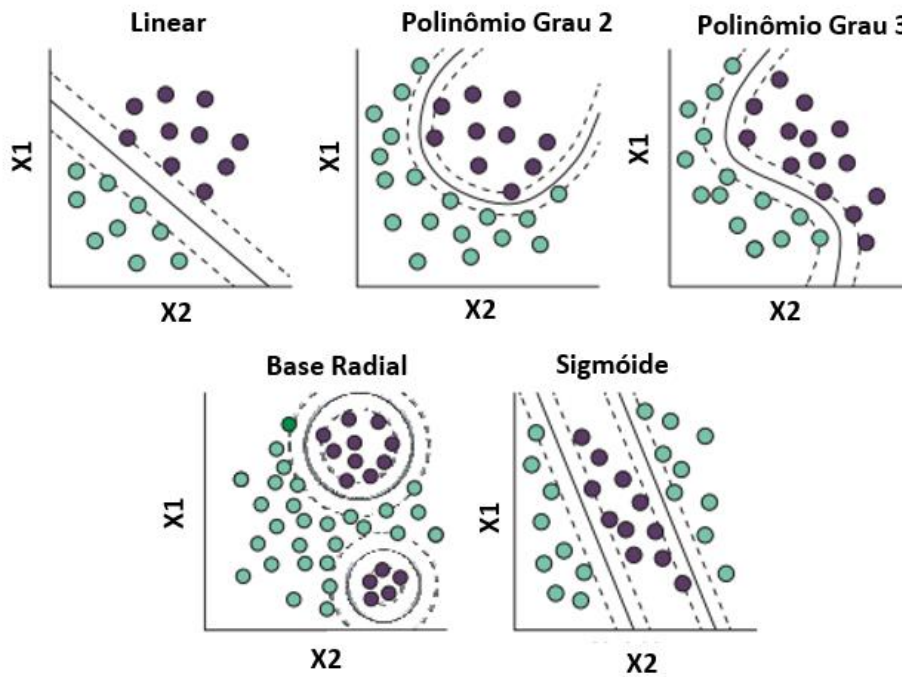
Fonte: RHYS (2020) modificado pelo autor

Existem vários tipos de funções Kernel que podem ser utilizadas no SVM, dentre os mais utilizados na prática pode-se citar: linear, polinomial, base radial (RBF) e sigmoide (**Equação 7**).

Pode-se observar na **Equação 7** que a função Kernel recebe dois vetores no espaço original e executa o seu produto interno no espaço de características. Na **Figura 9** ilustra-se exemplos de aplicação de vários tipos de funções Kernel. Para cada exemplo, a linha sólida indica o limite de decisão reprojetado no espaço de características original e as linhas tracejadas indicam as margens.

$$K(x, x') = \begin{cases} (x \cdot x') & \text{Linear} \\ (x \cdot x' + 1)^d & \text{Polinomial} \\ \exp(-\gamma \|x - x'\|^2) & \text{RBF} \\ \tanh(\gamma x \cdot x' + C) & \text{Sigmoidal} \end{cases} \quad \text{Equação 7}$$

Figura 9- Ilustração de exemplos de aplicação de vários tipos de funções Kernel. Para cada exemplo, a linha sólida indica o limite de decisão reprojeto no espaço de características original e as linhas tracejadas indicam as margens.



Fonte: RHYS (2020) modificado pelo autor

A técnica de utilização conjunta de vetores de suporte (SVM) com técnicas de regressão é chamada por alguns autores de SVR (*Support Vector Regression*). O SVR consiste em um algoritmo caracterizado pelo uso de um conjunto de funções: *kernel*, ε -função de intensidade de perda, e a função de controle obtida a partir do balanço de perdas e ganhos entre a maximização da margem e a suavidade da função $f(x)$. Para resolver um problema de regressão não-linear os dados de entrada são primeiramente transformados para o espaço de dimensão m , que pode ser maior que a dimensão original, onde os padrões de x são linearmente separáveis:

$$f(x) = \sum_{j=1}^m w_j \times \phi_j(x) + b \quad \text{Equação 8}$$

Onde w_j é um coeficiente de peso e b representa o viés, ϕ_j representa o conjunto de transformações não lineares no espaço de características. O objetivo do algoritmo SVR é estimar a função de regressão $f(x)$ que minimiza os seguintes termos:

$$C \frac{1}{n} \sum_{i=1}^n L_{\varepsilon}(f(x_i), y_i) + \frac{1}{2} \|w\|^2 \quad \text{Equação 9}$$

$$L_{\varepsilon}(f(x_i), y_i) = \begin{cases} 0, & |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, & |f(x) - y| \geq \varepsilon \end{cases} \quad \text{Equação 10}$$

As equações **Equação 9** e **Equação 10** descrevem a negociação entre o risco empírico e a total regularização. O termo $L_{\varepsilon}(f(x_i), y_i)$ é chamado de função de intensidade de perda, onde y representa os dados observados e ε representa a largura ou margem da zona de intensidade. O termo $C \frac{1}{n} \sum_{i=1}^n L_{\varepsilon}(f(x_i), y_i)$ é chamado de erro empírico que mede o desvio da amostra de treino fora das margens dos hiperplanos. C também é chamado de controle de capacidade que determina a negociação entre a tolerância ao erro empírico e a regularização do modelo dada pela parcela $\frac{1}{2} \|w\|^2$ (TAN; PERKOWSKI; PERKOWSKI, 2015).

O parâmetro C influencia diretamente a sensibilidade da função $f(x_i)$, quanto menor o valor de C , mais pobre será a complexidade da aproximação, e maior será a regularização e a especialização (*underfitting*) na amostra de treino. Por outro lado, maiores valores de C levam ao sobreajuste (*overfitting*) uma vez que no processo de otimização apresentado nas equações **Equação 9** e **Equação 10** o modelo dará maior peso à minimização do risco empírico (RAGHAVENDRA; DEKA, 2014).

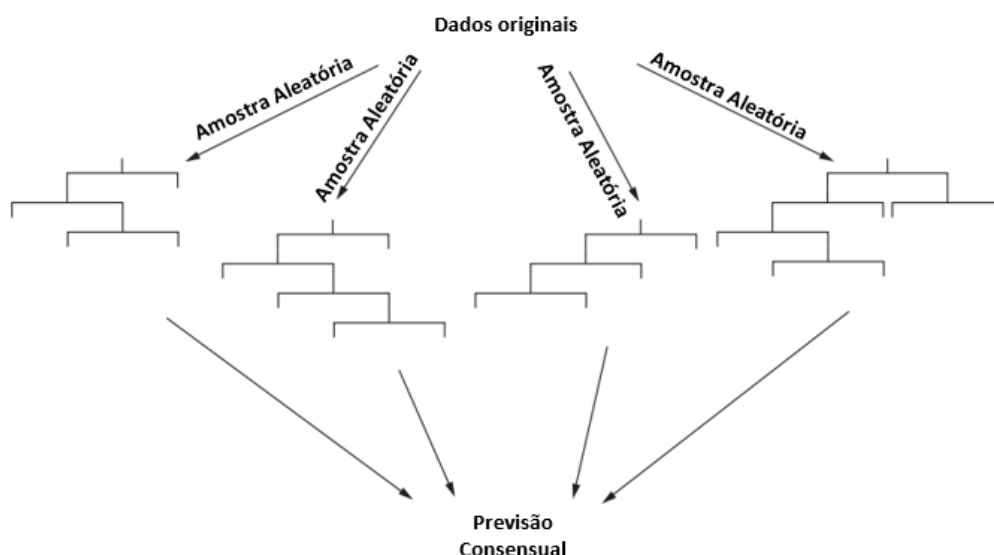
2.4.2 *Random Forest (RF)*

Árvores de decisão são modelos com muitas qualidades, dentre as quais pode-se enumerar: (1) São bastante intuitivos e simples; (2) são capazes de lidar com variáveis categóricas e contínuas; (3) não é necessário assumir nenhuma hipótese sobre a distribuição dos dados; (3) é capaz de lidar com falhas na aquisição de dados; (4) permite utilizar variáveis contínuas em diferentes escalas. Entretanto são fortemente suscetíveis ao sobreajuste (*overfitt*), tornando-se fortemente especializados na amostra de treino (RHYS, 2020).

Bootstrap Aggregation (Bagging) é um tipo de técnica *Ensemble* muito poderosa, que combina as previsões de múltiplos algoritmos *machine learning* juntamente para melhorar a acurácia das previsões individuais de um modelo. A técnica de *Bagging* é um procedimento que pode ser empregado com o objetivo de reduzir a variância de algoritmos que apresentam essa característica, como por exemplo as árvores de decisão (BROWNLEE, 2019).

A técnica de *Ensemble* por *Bagging* mais aplicada é o algoritmo *random forest* (BREIMAN, 2001) ilustrada na **Figura 10**. Esta técnica também apresenta boa performance de lidar com ruídos presente nos dados – algoritmos *machine learning* podem ser bastante sensíveis ao ruído presente nos dados resultados da presença de *outliers*, e este ruído uma vez presente na amostra de treino pode levar a modelos que apresentem previsões com elevada variância. No processo de amostragem com reposição (*bootstrapping*) há uma maior chance a selecionar dados próximos ao centro da sua distribuição ao invés dos extremos. Embora não haja nenhum impeditivo que um valor extremo (*outlier*) seja selecionado, a previsão será realizada com base em todos os modelos treinados. Simplesmente se deixa que eles votem qual o valor da previsão de acordo com a maioria. Esse recurso simplesmente reduz o efeito do ruído e reduz o risco de sobreajuste (*overfitting*) (RHYS, 2020).

Figura 10-Ilustração do mecanismo de funcionamento do algoritmo *random forest* (RF)



Fonte: RHYS (2020) modificado pelo autor

Entretanto o algoritmo *Random Forest* (RF) apresenta uma melhoria em relação a simplesmente *Bagging* uma vez que no processo de reamostragem a quantidade de preditores em cada amostra aleatória é restringida. Tipicamente se escolhe uma amostra de $m \approx \sqrt{p}$ do total de p preditores. Isso porque num processo de amostragem comum as árvores de decisão amostradas (*Bagged trees*) podem resultar em um conjunto bastante correlacionado, e desta forma, a média das previsões de modelos correlacionados não vai resultar em redução da variância. O algoritmo RF resolve esse problema forçando a escolha em um subconjunto de preditores (HASTIE; TIBSHIRANI; FRIEDMAN, 2017).

No campo da hidrologia RF já foi utilizado na previsão de áreas sujeitas a inundação (TEHRANY; PRADHAN; JEBUR, 2015), previsão de vazões diárias (PAPACHARALAMPOUS; TYRALIS, 2018) e previsão de secas (CHEN; LI; WANG, 2012), porém seu campo de aplicação mais frequente tem sido verificado no campo da classificação supervisionada de imagens multiespectrais em técnicas de sensoriamento remoto (BELGIU; DRĂGU, 2016).

Uma importante ferramenta, disponibilizada em decorrência da aplicação desta metodologia é o grau de importância do preditor, que pode ser calculado em função do quanto o erro decresce em cada ponto de separação (nó). A média da “queda” nos erros pode ser calculada como uma estimativa da importância da variável na formulação do modelo (BREIMAN, 2001; BROWNLEE, 2019).

2.4.3 Boosting XGboost

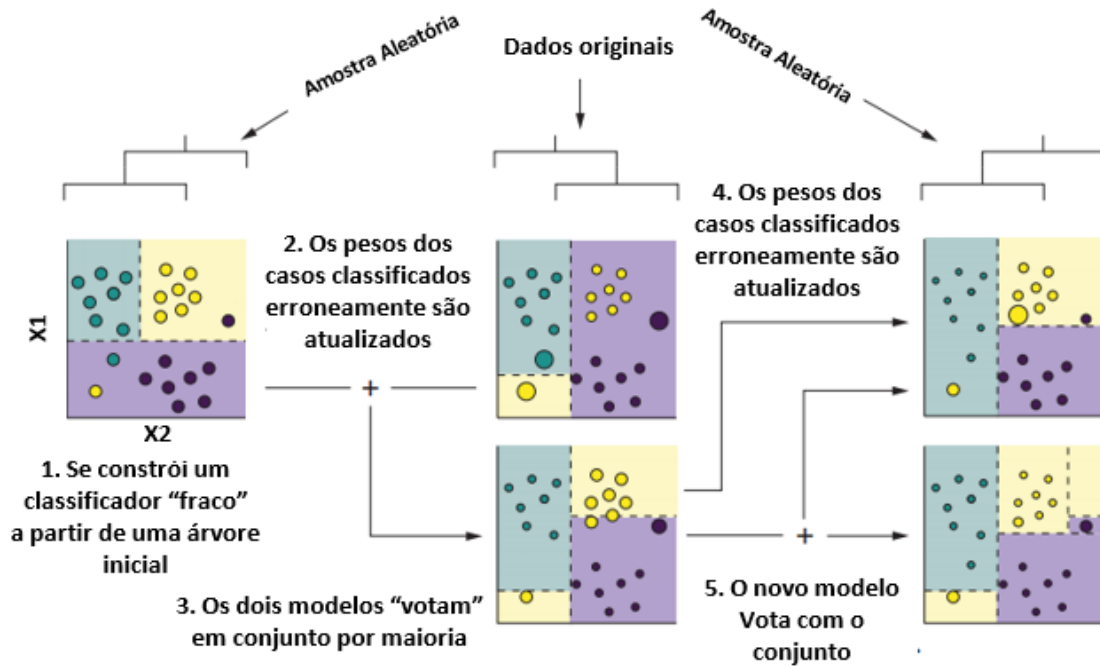
Boosting é um tipo de modelagem por conjunto “*ensemble*” que é empregada em alguns algoritmos, entre os mais conhecidos: Adaboost (FREUND; SCHAPIRE, 1996) que é um *boosting* adaptativo e XGBoost (CHEN; GUESTRIN, 2016) que é um *boosting* baseado no gradiente.

Diferentemente da técnica de *Bagging* em que os modelos são adicionados ao conjunto (*Ensemble*) em paralelo, na técnica de *Boosting* os modelos são adicionados individualmente em sequência, onde o modelo adicional busca corrigir os erros do conjunto de modelos previamente adicionados (RHYS, 2020). Um exemplo ilustrativo de funcionamento de um *boosting* de árvore de decisão utilizando o algoritmo Adaboost é aprestando na **Figura 11**.

O princípio da técnica de *Boosting* se baseia na adição sucessiva de classificadores fracos (*weak learners*) que são classificadores simples (não complexos) cuja acurácia é levemente superior a um processo aleatório ($p=50\%$). Assim o aumento da acurácia se dá pela adição de modelos, e não pela complexidade dos modelos que são adicionados. Quanto mais simples os modelos, maior o ganho de performance no treino RHYS (2020).

Gradient Boosting envolve três elementos: (1) uma função de perdas a ser otimizada; (2) um classificador fraco para se fazer previsões; (3) um modelo aditivo para minimizar o valor da função de perda. A função de perda deve ser diferenciável e nos problemas envolvendo regressão, o uso de funções quadráticas são recomendáveis (BROWNLEE, 2018).

Figura 11- Ilustração do mecanismo de funcionamento do algoritmo *Adaptive boosting with decision trees* (RF)



Fonte: RHYS (2020) modificado pelo autor

O algoritmo XGBoost (CHEN; GUESTRIN, 2016) é considerado o estado da arte da técnica de *Boosting*, durante as competições Kaggle no ano de 2015 esse algoritmo foi utilizado em 17 de soluções vencedoras do total de 29, entre essas 8 utilizaram unicamente essa técnica e as outras utilizaram de forma combinada com redes neurais. Para se ter uma comparação a técnica de aprendizado profundo foi utilizada em 11 soluções.

Neste algoritmo, o conjunto de árvores de decisão é adicionada ao conjunto (*ensemble*) de uma maneira aditiva até que o critério de parada seja atingido. Seu equacionamento matemático segundo (CHEN; GUESTRIN, 2016; NI et al., 2020) é apresentado a seguir nas Equação 11 a Equação 20,

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), \quad f_k \in F \quad \text{Equação 11}$$

Onde $\hat{y}_i$ é o valor previsto pelo modelo, f_k representa uma árvore de regressão individual pertencente ao conjunto de modelos, x_i são os vetores de entrada ou preditores, K é o número total de árvores de decisão pertencentes ao modelo e F é o conjunto de todas as f_k s. O processo de aprendizado envolve a minimização da função objetivo regularizada $L(\Phi)$ a seguir

$$L(\emptyset) = \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(f_k)$$

Equação 12

$$\text{onde } \Omega(f_k) = \gamma \cdot T + \frac{1}{2} \cdot \lambda \cdot \|\omega\|^2$$

Onde l , é uma função de perda entre o valor observado (y_i) e o previsto ($\hat{y}_i$) pela função. O termo (Ω) da equação é o termo de regularização que penaliza a complexidade do modelo para evitar o risco de sobreajuste e γ é a complexidade de cada folha da árvore. T é quantidade de folhas de cada árvore. λ é o peso que é dado na escala de penalidade. ω é o vetor de escores.

O modelo é treinado de forma aditiva. Seja $\hat{y}_i^t$ a previsão da i -ésima instância no t -ésima iteração, uma nova função f_t é adicionada para minimizar a seguinte função objetivo

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t)$$

Equação 13

Para simplificar a otimização da Equação 13, a função de perda é expandida por uma série de Taylor de segunda ordem,

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(a)}{2!}(x - a)^2$$

Equação 14

Tomando-se $x=L^{(t)}$ e $a=\hat{y}_i^{t-1}$ a função objetivo pode ser rapidamente otimizada por

$$L^{(t)} \cong \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Equação 15

Onde $g_i = \partial_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{t-1})$, e $h_i = \partial_{\hat{y}_i^{t-1}}^2 l(y_i, \hat{y}_i^{t-1})$ são os gradientes de primeira e segunda ordem da função de perda, respectivamente. Retirando o termo constante da Equação 15 para simplificar a função objetivo no tempo t , tem-se

$$\tilde{L}^{(t)} = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Equação 16

Definindo-se I_j como a j -ésima folha de instância de treinamento de uma estrutura $q(x_i)$, e expandindo-se o termo (Ω) a **Equação 16** pode ser reescrita como a seguir

$$\begin{aligned}\tilde{L}^{(t)} &= \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma \cdot T + \frac{1}{2} \cdot \lambda \cdot \sum_{j=1}^T \omega_j^2 \\ &= \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) \cdot \omega_j + \frac{1}{2} \cdot \left(\sum_{i \in I_j} h_i + \lambda \right) \cdot \omega_j^2 \right] + \gamma \cdot T\end{aligned}\tag{Equação 17}$$

Para uma estrutura $q(x_i)$, os pesos ótimos ω_j^* de uma j -ésima folha é dada por

$$\omega_j^* = \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} + \gamma \cdot T\tag{Equação 18}$$

A Equação 18 pode ser usada como uma função de *scores* para mensurar a q qualidade da estrutura de uma árvore de decisão $q(x_i)$. Esse escore pode ser interpretado como uma estimativa da impureza da árvore de decisão.

$$\tilde{L}^{(t)} = -\frac{1}{2} \cdot \sum_{j=1}^T \left[\frac{\left(\sum_{i \in I_j} g_i \right) \cdot \omega_j}{\sum_{i \in I_j} h_i + \lambda} \right] + \gamma \cdot T\tag{Equação 19}$$

Assumindo-se que I_L e I_R são as instâncias de treinamento antes e depois da divisão, a redução da perda após a divisão é dada por

$$L_{split} = \frac{1}{2} \cdot \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} + \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma\tag{Equação 20}$$

Essa fórmula é utilizada na prática para selecionar os melhores candidatos.

A técnica ainda foi pouco aplicada no campo da hidrologia, entre eles podem-se destacar previsão de risco de inundação (MA et al., 2021) e simulação de vazões (JI et al., 2021).

2.5 TRANSFORMAÇÃO DE DADOS

Transformações de dados em ML são operações realizadas principalmente sobre os preditores com os mais variados objetivos, dentre os quais pode-se citar: reduzir ou eliminar características indesejáveis como a colinearidade, ausência de normalidade e assimetria entre os dados; reduzir dimensões de grandes massas de dados, eliminar ruídos, remover o efeito do valor numérico da variável durante a modelagem, dentre outros.

Transformações de padronização centram os dados em torno da média e padronizam os dados $\frac{x-\mu}{\sigma}$. Transformações de normalização escalam os dados dentro de um intervalo, sendo mais comum [0,1]. As transformações do tipo Box-Cox (BOX; D.R. COX., 1964), log-normal (STOLINE, 1991) e Yeo-Johnson (YEO; JOHNSON, 2000) são usadas para reduzir a assimetria e curtose buscando a normalidade dos dados.

As transformações de redução de dimensionalidade são geralmente rotações aplicadas para se trabalhar com um número menor possível de dimensões (variáveis) de forma que estas ainda retenham o máximo das informações do conjunto original. A redução de dimensões potencializa a visualização de relações entre variáveis, identificação de padrões e estruturas presentes nos dados (MARTINEZ; MARTINEZ, 2004; RHYS, 2020).

Dentre os métodos mais empregados para redução de dimensões pode-se citar: a análise de componentes principais (PCA); decomposição do valor singular (SVD) e análise fatorial (FA). Todas essas técnicas se tratam de rotações e projeções aplicadas aos dados em sua forma matricial, e muitas vezes são tratadas como sinônimos devido à grande similaridade da fundamentação matemática que sustenta todas as abordagens: a rotação por transformação linear. Entretanto, conceitualmente são abordagens distintas.

Através da análise das PCAs consegue-se explicar a variância através de k combinações lineares das variáveis originais $\mathbf{p}$, sendo que $k < p$, e que grande parte da variabilidade do sistema é explicada pelas k componentes. As k componentes podem então substituir as p variáveis originais.

As componentes principais são expressas como combinações lineares das variáveis originais $X = [X_1, X_2, \dots, X_p]$ sendo dependentes da matriz de covariâncias S ou de correlações R .

A seguir a formulação algébrica da PCA dada por (FERREIRA, 2000): sendo S uma matriz de covariâncias associada ao vetor $X = [X_1, X_2, \dots, X_p]$ e os pares de

autovalores e autovetores $((\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p))$ onde $\lambda_1 \geq \lambda_2 \dots \geq \lambda_p$ são todos não nulos, a componente principal de ordem i é dada por:

$$Y_i = \mathbf{e}_i' \cdot \mathbf{X} = \mathbf{e}_{1i} \cdot X_1 + \mathbf{e}_{2i} \cdot X_2 + \dots + \mathbf{e}_{pi} \cdot X_p \text{ onde } i = 1, 2, \dots, p$$

As componentes principais não são correlacionadas:

$$[cor(Y_i, Y_k) = \mathbf{e}_i' \cdot \mathbf{e}_k = 0, \text{ para } i \neq k]$$

E tem variâncias iguais aos autovalores de $\mathbf{S}$:

$$[var(Y_i) = \mathbf{e}_i' \cdot \mathbf{e}_i = \lambda_i, \text{ para } i = 1, 2, \dots, p]$$

Outras relações importantes decorrem da obtenção das componentes principais:

A variância original do conjunto de variáveis não é modificada após a transformação,

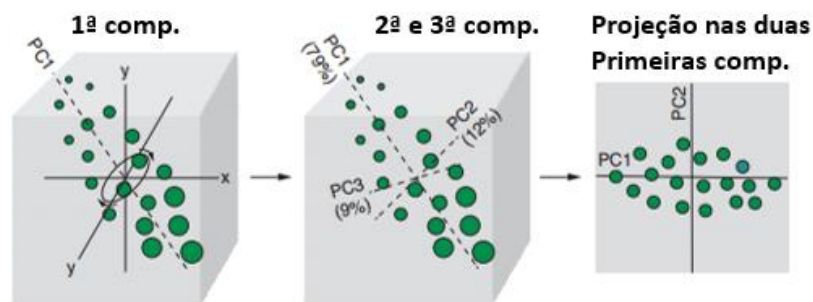
$$\sigma_{11} + \sigma_{22} + \dots + \sigma_{pp} = \sum_{j=1}^p Var(X_j) = \lambda_{11} + \lambda_{22} + \dots + \lambda_{pp} = \sum_{j=1}^p Var(Y_j)$$

A proporção da variância explicada por cada componente é dada por,

$$\left(\begin{array}{c} \text{Proporção da} \\ \text{Variância explicada} \\ \text{Pela componente } k \end{array} \right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p}, k = 1, 2, \dots, p$$

A **Figura 12** ilustra o conceito das PCAs em um espaço tridimensional, a primeira componente é posicionada na direção da maior variabilidade dos dados, em sequência a segunda componente é posicionada ortogonal a primeira, mas ainda existe uma liberdade de rotação representada pela elipse, que é fixada quando se maximiza a variância remanescente da primeira PC. A terceira componente é ortogonal às duas primeiras e não possui liberdade de rotação, o que explica a baixa quantidade de variância.

Figura 12-Ilustração de funcionamento de uma PC em um espaço tridimensional.



Fonte: (RHYS, 2020)

Uma técnica também muito importante, que partilha da mesma fundamentação matemática das PCAs é a análise fatorial (FA), e devido a isso muitos textos da literatura não fazem distinção entre ambas.

Na verdade, ambas são técnicas de redução de dimensionalidade empregadas com o objetivo de substituir um largo conjunto de variáveis por um conjunto menor. Mas em uma abordagem simplificada, se utiliza a PCA quando se necessita simplificar seus dados ou ter uma compreensão deles utilizando menos dimensões (como visualização por exemplo). Já a análise Fatorial (FA) é utilizada quando se necessita de um modelo explanatório que explique as correlações entre os seus dados (MATHWORKS, 2004).

Uma diferença entre os dois métodos está no conceito, enquanto na PCA se pretende rotacionar as variáveis projetando-as em um novo sistema de coordenadas rotacionado e ortogonal, na FA se deseja reescrever as variáveis como combinação de $d < p$ fatores f_j como a seguir

$$X_1 = \lambda_{11} \cdot f_1 + \dots + \lambda_{1d} \cdot f_d - \varepsilon_1$$

....

$$X_p = \lambda_{p1} \cdot f_1 + \dots + \lambda_{pd} \cdot f_d - \varepsilon_p$$

As variáveis λ_{ij} para $(i=1..p \text{ e } j=1..d)$ são denominadas *Cargas nos fatores* e os termos ε_i são denominados *Especificidades dos fatores*, sendo específico para cada uma das variáveis originais, enquanto o termo f_j são comuns a todas as variáveis. A soma dos quadrados das cargas nos fatores é chamada de comunalidade da variável X.

No campo das aplicações em meteorologia e das geociências as análises de fatores FA tem sido utilizadas largamente para decompor campos escalares de variáveis

espaço-temporais (x,y,t) em padrões espaciais (x,y) e sinais temporais (t) . Nestas aplicações, recebem o nome de *Funções Empíricas Ortogonais*- mais conhecidas como EOFs. Sua formulação matemática apresentada a seguir se fundamenta em MENKE & MENKE (2012) e DEWITTE et al. (2011)

$$f(x, y, t) = \sum_{n=1}^N f_n(x, y) \cdot f_n(t)$$

onde $f_n(x, y)$ são os padrões modais e $f_n(t)$ são os fatores.

No caso da FA a rotação nos fatores pode ser ortogonal ou oblíqua, diferentemente da PCA. Diversos métodos de rotação podem ser aplicados, a saber: *quartimax*, *varimax*, *orthomax* e *equimax* (ortogonais) e *promax* e *procrustes* (oblíquos) (MARTINEZ; MARTINEZ, 2004).

HANNACHI; JOLLIFFE; STEPHENSON (2007) comentam que as restrições de ortogonalidade e geração de variáveis temporais não correlacionadas das EOFs e PCAs, impõem severos limites na interpretação do significado físico dos padrões espaciais gerados. Isso acontece porque os processos físicos não são independentes e, portanto, espera-se que os modos não sejam necessariamente ortogonais. Por este motivo, os métodos citados acima são empregados para aliviar as restrições, e assim produzir estruturas espaciais mais simples e de mais fácil interpretação.

2.6 INDICES DE SECA

2.6.1 *Índice de Precipitação Padronizada (SPI)*

Partindo do princípio de que os déficits de precipitação se conectam à diversas reservas utilizáveis de água (umidade do solo, reservas subterrâneas, águas superficiais, estoques de neve e barragens) em diferentes escalas temporais. E, conseqüentemente, seu impacto depende da duração em que os déficits de precipitação são acumulados. Visando suprir tomadores de decisão de um índice versátil para monitorar e analisar os impactos de secas no Colorado (EUA), MCKEE; DOESKEN; KLEIST (1993) desenvolveram o Índice de Precipitação Padronizada (SPI) tornando possível expressá-las em termos de probabilidade, déficits e suprimento de água – em diferentes escalas temporais.

Uma das grandes vantagens do SPI é que seu cálculo requer somente o dado de precipitação associado a uma escala temporal e segundo os autores pode ser aplicado de forma similar para outras variáveis hidrológicas e climáticas, a saber: vazões, armazenamento em reservatórios, estoques de neve, umidade do solo.

Para sua estimativa, os autores prescrevem os seguintes passos:

1. Prepara-se uma série temporal de dados contínua de precipitações mensais de no mínimo 30 anos;
2. Adotar uma escala temporal arbitrária i em meses para os quais serão calculadas as precipitações médias. Normalmente $i = 3, 6, 12, 24, 48$ meses, são escalas temporais típicas em que os déficits de precipitação afetam os diversos tipos de reservas utilizáveis de água referidas acima;
3. Ajusta-se a função densidade de probabilidade *Gamma*, apresentada nas **Equação 21** e **Equação 22** (KUSHLER, 2002), aos valores médios de precipitação calculados no passo anterior, a fim de serem calculadas as respectivas probabilidades de ocorrência;
4. Com a distribuição de probabilidades acumuladas, correspondentes a cada valor de precipitação média estimada no passo 2, entra-se na distribuição normal padronizada e calcula-se de forma inversa o valor da variável padronizada, sendo uma medida dos desvios em relação à média zero com desvio padrão unitário;
5. O valor obtido no passo 4, através da distribuição normal padronizada é o SPI.

$$f(x, \lambda, t) = \frac{\lambda e^{-\lambda x} (\lambda x)^{t-1}}{\Gamma(t)} ; x \geq 0, \quad \text{Equação 21}$$

$$\Gamma(t) = \int_0^{\infty} e^{-y} y^{t-1} dy \quad \text{Equação 22}$$

Um evento de seca é definido como um período no qual o SPI é continuamente negativo com valores menores ou iguais a -1. A seca se inicia com valores de SPI menores que zero e termina quando os valores de SPI se tornam maiores que zero.

A fim de caracterizar a intensidade das secas, os autores propuseram os seguintes intervalos de classificação apresentados na **Tabela 2**.

Tabela 2-Classificação de intensidade de secas segundo o índice SPI

Valores de SPI	Classificação de Secas
0 a -0,99	Seca leve
-1,00 a -1,49	Seca Moderada
-1,50 a -1,99	Seca Severa
$\leq -2,00$	Seca Extrema

Fonte: (MCKEE; DOESKEN; KLEIST, 1993)

O SPI é um índice de natureza meteorológica, considerado na avaliação de KEYANTASH; DRACUP (2002) um dos mais robustos, transparentes e de fácil obtenção, dentre outras qualidades. Suas aplicações têm sido dentre as mais diversas: (1) monitoramento de secas hidrológicas (MELO et al., 2016); previsão de secas meteorológicas (PAULO; PEREIRA, 2007; BELAYNEH et al., 2014b; BELAYNEH; ADAMOWSKI; KHALIL, 2016); previsão de secas hidrológicas (NALBANTIS; TSAKIRIS, 2009); investigação de tele-conexões climáticas (COSTA et al., 2016).

Não obstante o seu largo campo de aplicabilidade, o seu emprego na previsão de vazões não é encontrado na literatura. Alguns fatores podem contribuir para que isso aconteça: (1) embora a precipitação seja correlacionada com a vazão, o processo de cálculo do índice passa por duas transformações de dados a fim de padronizá-lo, o que contribui com o enfraquecimento das relações diretas com o valor numérico da vazão, mas não das anomalias de vazão que são tratadas pelos índices de seca; (2) o índice é fortemente associado a uma escala temporal, e esta é uma decisão a ser arbitrada antes do seu computo. Embora seja consenso na literatura de que escalas de 12 e 24 sejam suficientes para representar os processos hidrológicos, o processo de previsão de vazões necessita de maior precisão da estimativa dessa janela temporal.

Além disso, a defasagem temporal entre a resposta do sistema hidrológico às mudanças meteorológicas medidas por este índice em rios não regularizados chega a aproximadamente 1 mês (~6 meses nos estoques dos reservatórios), conforme análises conduzidas por MELO; WENDLAND (2016). Antecedência (habilidade) considerada pequena para os propósitos deste trabalho.

2.6.2 Índice de Severidade de Seca Palmer (PDSI)

Este índice foi desenvolvido por Palmer (PALMER, 1965) e já no início dos anos 80 já era utilizado em boletins agro-climatológicos oficiais como a NOAA para monitoramento de secas nos EUA (ALLEY, 1984) e por profissionais e empreendedores do agronegócio com a finalidade de monitorar tendências hidrológicas, previsão de

produtividade de culturas, monitoramento de risco de incêndios (HEDDINGHAUS; SOBOL, 1991).

Trata-se de um índice baseado em um balanço hídrico entre suprimento e retiradas de umidade em duas camadas de solo, existindo três variações deste indicador: meteorológico (PDSI), hidrológico (PHDI) e agro-meteorológico (Z-index) (WEBER; NKEMDIRIM, 1998). Essas variações dizem respeito principalmente aos mecanismos e critérios que sua formulação considera para início ou término de um período seco ou úmido.

Segundo a avaliação de KEYANTASH; DRACUP (2002) trata-se de um índice sofisticado, de difícil operacionalidade, baixa transparência, e pouca robustez. Ao longo da sua criação o índice Palmer sofreu duras críticas, especialmente por ALLEY (1984) e KARL (1986), dentre as quais pode-se citar: (1) elevada sensibilidade ao período de calibração; (2) a relação entre a severidade e duração (**Tabela 3**) foi obtida de forma empírica e gráfica para regiões central do Iowa e oeste do Kansas; (3) o coeficiente K que ajusta os desvios de suprimento de umidade (precipitação) em relação à climatologia local são dependentes do local onde o índice foi inicialmente aplicado; (4) é um índice extremamente auto correlacionado em decorrência da sua formulação; (5) A espessura das camadas de solo foram arbitrárias; (6) Não existe nenhum atraso (“lag”) no balanço hídrico realizado pelo modelo. ALLEY (1984) ainda discute problemas de utilização desse índice em modelos autorregressivos devido à forte autocorrelação temporal imposta na sua formulação

$$X(t) = \frac{Z(t)}{3} + c \cdot X(t - 1) \quad \text{Equação 23}$$

$$Z(t) = K \cdot d \quad \text{Equação 24}$$

$$d = P - (\alpha PE + \beta PR + \gamma PRO - \delta PL) \quad \text{Equação 25}$$

$X(t)$ é o índice PDSI; $Z(t)$ é a índice previamente à suavização, considerado o Palmer agrometeorológico cujas variações são rápidas em função da quantidade de precipitação; c é o índice de autocorrelação, ajustado por Palmer inicialmente como sendo $c=0,897$ e d são os desvios em relação à climatologia local. O termo à direita da Equação 25 é a precipitação CAFEC (*Climatically Appropriate for Existing Conditions*). A proposta do fator K é ajustar o valor de “ d ” de acordo com as características do clima de forma a permitir comparações do índice em tempos e espaços distintos.

Tabela 3-Classificação de intensidade de secas - PDSI

Valores de PDSI	Classificação de Secas
$\geq 4,0$	Extremamente úmido
3,0 a 3,99	Muito úmido
2,00 a 2,99	Moderadamente úmido
1,00 a 1,99	Levemente úmido
0,50 a 0,99	Incipiente úmido
0,49 a -0,49	Normal
-0,50 a -0,99	Seca incipiente
-1,00 a -1,99	Levemente seco
-2,00 a -2,99	Moderadamente seco
-3,00 a -3,99	Muito seco
$\leq -4,00$	Extremamente seco

Fonte: ALLEY (1984)

O índice Palmer resistiu às duras críticas sofridas nos anos 80, sendo revisitado em WEBER; NKEMDIRIM, (1998) quando são enaltecidas todas as suas qualidades necessárias a função de um índice de seca. Este trabalho inclusive ressalta a habilidade do PDSI em se antecipar 1 mês mudanças na umidade do solo. E a resposta do PDSI em relação ao PHDI em 90-120 dias.

O problema da representatividade espacial foi resolvido por WELLS; GODDARD; HAYES, (2004) trinta anos após sua criação. Uma auto calibração do PDSI denominada scPDSI foi apresentada pelos autores, ajustando as constantes do índice de forma dinâmica para as localidades de aplicação.

Desde então, o scPDSI tem sido disponibilizado em bases globais (DAI; TRENBERTH; QIAN, 2004, VAN DER SCHRIER et al., 2013) e tem sido usado em avaliações de tendências climáticas globais (DAI, 2011a), mostrando correlações bastante significativas entre o índice e umidade do solo e deflúvios superficiais em grandes e pequenas bacias em várias partes do mundo.

2.6.3 NOAA - AVHRR Vegetation Condition Index (VHI)

O VHI é um índice disponibilizado pela NOAA desde 1981, sendo calculado a partir dos dados adquiridos pelo sensor *Advanced High Resolution Radiometer* (AVHRR), sendo uma composição dos índices *Vegetation Condition Index* (VCI) e *Temperature Condition Index* (TCI).

O VCI é calculado a partir do NDVI, que por se tratar de um índice que responde positivamente à vegetação e negativamente a solos descobertos, nuvens e água

foi a primeira aproximação na tentativa de monitoramento das secas. Entretanto, observou-se que em alguns casos, somente a utilização do VCI isoladamente não apresentou acurácia satisfatória na detecção e análise de secas. De forma que as bandas termais do sensor foram adicionadas buscando-se agregar mais informação ao índice, e a partir daí o TCI foi desenvolvido (KOGAN, 1995).

Para o cálculo do VCI, o algoritmo desenvolvido por KOGAN (1995) separa o sinal do NDVI em duas componentes: a) alta-frequência que é relacionada a flutuações das condições atmosféricas e b) baixa-frequência que é relacionada a flutuações de longo-termo do ambiente e ecossistemas. O filtro apresentado na equação a seguir é aplicado então sobre o sinal do NDVI de forma a obter o VCI

$$VCI = 100 \times \frac{(NDVI - NDVI_{min})}{(NDVI_{max} - NDVI_{min})} \quad \text{Equação 26}$$

Onde o NDVI, $NDVI_{max}$ e $NDVI_{min}$ são os sinais suavizados semanais do NDVI, máximo e mínimo da série, respectivamente.

De forma semelhante o TCI segue a expressão

$$TCI = 100 \times \frac{(T_{max} - T)}{(T_{max} - T_{min})} \quad \text{Equação 27}$$

De forma oposta ao NDVI, elevadas temperaturas no meio de uma estação indicam condições desfavoráveis (ou seca) e baixas temperaturas condições favoráveis.

O VHI se trata de uma composição dos dois índices, de acordo com KOGAN (1997) a sua formulação é dada por

$$VHI = \alpha \times VCI + (1 - \alpha) \times TCI \quad \text{Equação 28}$$

onde α é geralmente adotado como 0,5.

A utilização dos índices VCI-TCI é uma sólida fonte de informação sobre o stress hídrico da vegetação, a qual a seca é a maior causa. Ainda, pode ser considerada bastante útil como uma forma de diagnóstico em tempo real das condições da vegetação especialmente se dados meteorológicos de monitoramento não estiverem disponíveis. E mesmo que existam sua utilização de forma consorciada e combinada com os índices derivados do AVHRR podem trazer ganhos no monitoramento de estresse da vegetação,

estimativas de secas e impactos meteorológicos. Este índice também tem sido testado e empregado com sucesso no monitoramento dos impactos provocados pela seca na produção agrícola em todo o mundo (KOGAN, 1997), e mostrado também habilidade em antecipar eventos de seca (KOGAN; GUO, 2016), tanto em larga como pequena escala (KOGAN, 1995).

Com base nos trabalhos de KOGAN (2002) e MCKEE; DOESKEN; KLEIST (1993) de BRITO et al. (2018) relacionam os dois índices na **Tabela 4**.

Tabela 4-Classificação de secas relacionando os índices SPI e VHI

Classificação de Secas	Valores de SPI	VHI
Seca leve	$-1,0 < \text{SPI} < 1,0$	$\text{VHI} > 40$
Seca Moderada	$-1,5 < \text{SPI} \leq 1,0$	$20 < \text{VHI} \leq 30$
Seca Severa	$-2,0 < \text{SPI} \leq -1,5$	$10 < \text{VHI} \leq 20$
Seca Extrema	$\text{SPI} \leq -2,0$	$\text{VHI} \leq 10$

Fonte: BRITO et al. (2018)

Este índice tem mostrado correlações significativas com o SPI, cujo grau de associação varia de acordo com a escala temporal utilizada no cálculo da anomalia de precipitação, sendo maior nas escalas de 3 a 9 meses (DUTTA et al., 2015).

3 MATERIAIS E MÉTODOS

3.1 DESCRIÇÃO GERAL

Para avaliar a qualidade de previsões sazonais em período de estiagem fazendo uso de algoritmos de aprendizado estatístico (*machine learning* - *ML*) a partir de produtos derivados de sensoriamento remoto disponibilizados em escala global foi aplicada a metodologia apresentada na **Figura 13**.

Primeiro os dados de índices de seca e precipitação, disponibilizados em formato raster e georreferenciados no sistema WGS84, foram recortados a partir dos limites das bacias em formato *shapefile*. Parte dessa operação foi realizada no Google Earth Engine - GEE (GORELICK et al., 2017) para os índices Palmer (PDSI) e precipitação (CHIRPS).

O processamento da extração das imagens referente aos índices de seca *Vegetation Health Index* - *VCI* e *Temperature Condition Index* – *TCI*, disponibilizados em formato netcdf, foi realizada de modo *off-line* após seu *download*. Esse processo foi realizado com o uso do software RStudio (RSTUDIO TEAM, 2022) e das bibliotecas **raster** (HIJMANS, 2020), **sp** (BIVAND; PEBESMA; GOMEZ-RUBIO, 2013), **sf** (EDZER PEBESMA, 2018).

Após extração, os dados espaciais foram convertidos em formato tabular em formato *dataframe*, para serem emparelhados com as vazões observadas e convertidos em séries temporais.

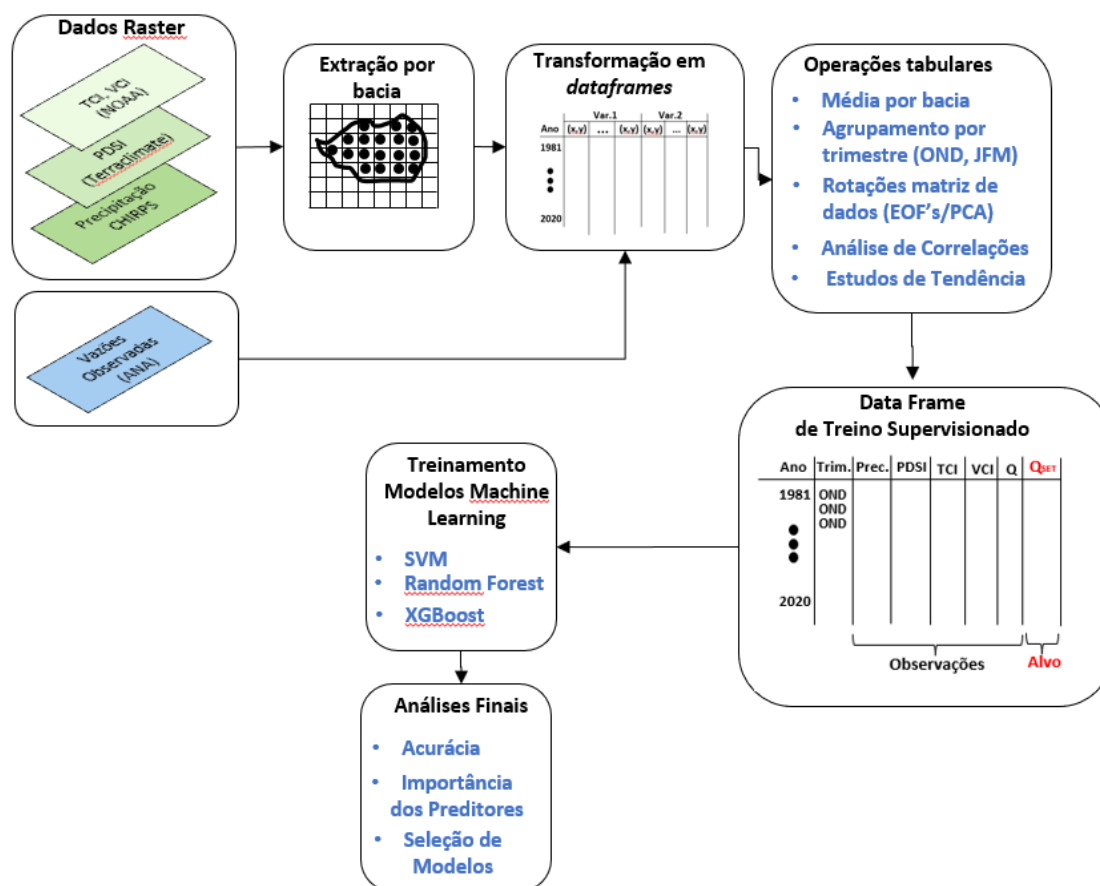
Os *dataframes* foram construídos tanto sobre valores médios dos pixels inscritos no polígono da bacia, como também mantendo individualmente o valor de cada pixel para posterior análise de distribuição espacial e extração das funções empíricas ortogonais (EOFs).

As previsões foram realizadas no meio da estação chuvosa (após o término do primeiro trimestre do ano hidrológico) e no término da estação chuvosa (após o segundo trimestre), portanto com 9 e 6 meses de antecedência, considerando o ano hidrológico se iniciando em outubro de cada ano e com término em setembro do ano seguinte.

Desta forma os dados foram agrupados por trimestre OND e por semestre ONDJFM. Esse processo foi realizado utilizando manipulação de dados no software RStudio e com as bibliotecas **dplyr** (WICKHAM et al., 2021) e **reshape2** (WICKHAM, 2007).

Após serem analisados e convertidos em formato tabular, condição necessária de entrada no modelo, o treinamento dos algoritmos *machine learning* para obtenção dos modelos de previsão foram realizados no Rstudio com a biblioteca **caret** (KUHN, 2020).

Figura 13-Fluxograma geral da metodologia



Fonte: Elaborado pelo autor

3.2 ÁREA DE ESTUDO

Para realização dos experimentos foram selecionadas 5 bacias hidrográficas localizadas entre as latitudes $-20,6^\circ$ e $-13,4^\circ$ e longitudes $-46,7^\circ$ e $-39,0^\circ$, e que atendessem preliminarmente aos critérios: (a) apresentassem baixo comprometimento hídrico durante o período de estiagem, com o objetivo de evitar treinar o modelo sobre dados não representativos das condições naturais e sem correspondência às variáveis forçantes adquiridas por meio de sensoriamento remoto; (b) representassem um conjunto diverso

quanto ao regime pluviométrico e variabilidade climática, permitindo assim avaliar o comportamento dos modelos estatísticos ML treinados em situações distintas e (c) apresentassem características não estacionárias nos sinais climatológicos e hidrológicos para o período estudado.

Dentre as bacias selecionadas o Alto São Francisco (ALTO_SF), Alto Jequitinhonha (ALTO_JQ) e Vacaria (VACARIA), se localizam no bioma Cerrado com período chuvoso iniciando em outubro. As bacias dos Ribeirão São Francisco (Rib_SF) e Gongogi (Gongogi) fazem parte do bioma Mata Atlântica.

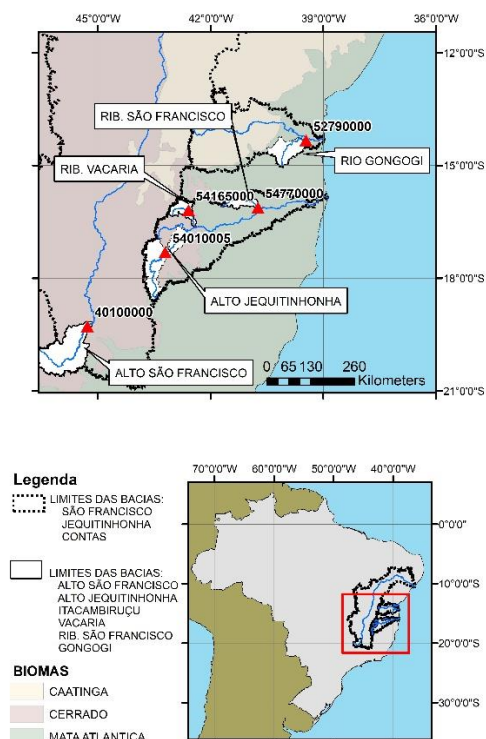
Segundo a classificação de Köppen para o território brasileiro (ALVARES et al., 2013) a área selecionada abrange 06 tipos climáticos, conforme **Figura 15** e **Tabela 5**. São eles: subtropical úmido com chuvas de verão quentes (Cwa) e temperados (Cwb) típicos do bioma cerrado; tropical com chuvas de inverno (As) e com chuvas de verão (Aw) mais presentes nas bacias na transição da mata atlântica-cerrado. Cabe lembrar que no hemisfério sul o verão é definido pelo período dos seis meses mais quentes ONDJFM e o inverno período correspondente aos seis meses mais frios AMJJAS.

A ocorrência de clima subtropical com verões temperados (Cwb) identificados na área de estudo estão relacionados a bacias no bioma cerrado associados a presença de elevadas altitudes (bacias ALTO_JQ e ALTO_SF), conforme **Figura 14** e **Figura 15**. As áreas das bacias selecionadas variam de aproximadamente 3.000 a 14.000 km² conforme **Tabela 5**.

Nas **Figura 17** e **Figura 18** são apresentadas as climatologias de precipitação e escoamentos anuais em mm·mês⁻¹, onde observa-se a ocorrência do período chuvoso no semestre de ONDJFM e do período seco em AMJJAS, sendo setembro o mês mais seco em relação ao aspecto hidrológico.

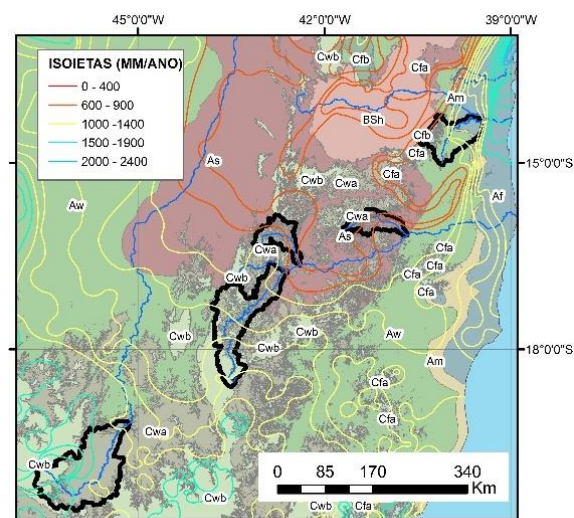
Em termos de aspectos hidrogeológicos (**Figura 16**) as bacias dos rios Vacaria, Rib_SF e Gongogi apresentam uma recarga anual da ordem de 20-100 mm·ano⁻¹, sendo caracterizadas por vulnerabilidade hidrogeológica moderada para cheias e secas, e uma pequena área das cabeceiras da bacia do rio Gongogi se encontra em uma região de baixa recarga e vulnerabilidade alta para secas. As bacias dos rios ALTO_SF e ALTO_JQ se caracterizam por apresentar melhores condições de recarga (>100 mm/ano) e vulnerabilidade alta para cheias e moderada para secas.

Figura 14-Localização da área de estudo (Elaborado a partir de informações disponibilizadas no SNIRH/ANA, IBGE e MMA).



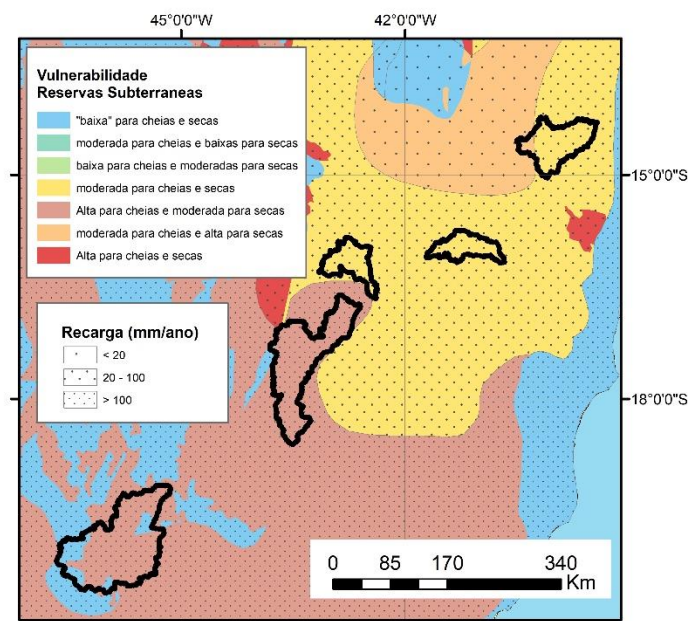
Fonte: Elaborado pelo autor

Figura 15-Isoietas de precipitação anual elaborada com dados no período de 1977-2006 (CPRM; ANA, 2016) produzidos a partir de dados da rede hidrometeorológica nacional e classificação climática de Köppen elaborado a partir de



Fonte: Elaborado pelo autor a partir de (ALVARES et al., 2013)

Figura 16-Recarga anual e vulnerabilidade das reservas subterrâneas



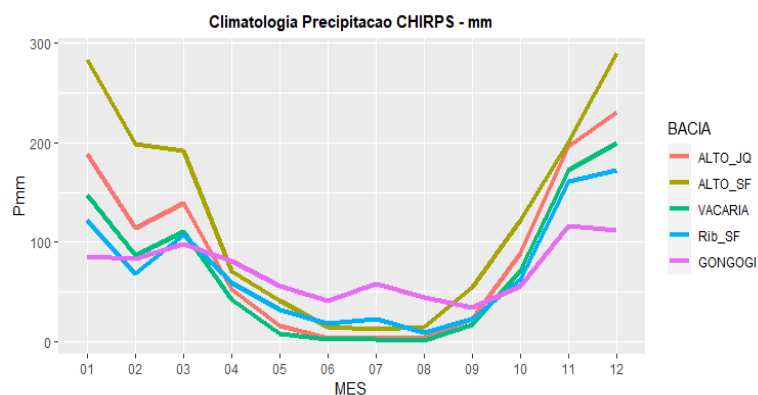
Fonte: RICHTS; STRUCKMEIER; ZAEPEKE, (2011) modificado pelo autor

Tabela 5-Características físicas e climáticas das bacias dos rios selecionados para realização da previsão.

BACIA	Rio	Área (km ²)	Período Seco (Chuvoso)	Precipitação anual (mm)	Altitude (m) (min/méd/máx.)	Clima predominante (Köppen)/Bioma
Alto SF	São Francisco	14.147	abr-set (out-mar)	1.494	573/784/1.509	Cwa/Cwb (Cerrado)
Alto JQ	Jequitinhonha	9.509	abr-set (out-mar)	1.058	345/878/2.058	As/Cwa/Cwb (Cerrado)
Vacaria	Vacaria	3.195	abr-set (out-mar)	862	305/815/1.367	As/Cwa (Cerrado)
RibSF	Rib. São Francisco	2.899	abr-set (out-mar)	857	175/716/1.075	As/Cwa (Mata Atlântica)
Gongogi	Gongogi	5.336	abr-set (out-mar)	866	43/473/1.187	Af/Aw/Bsh (Mata Atlântica)

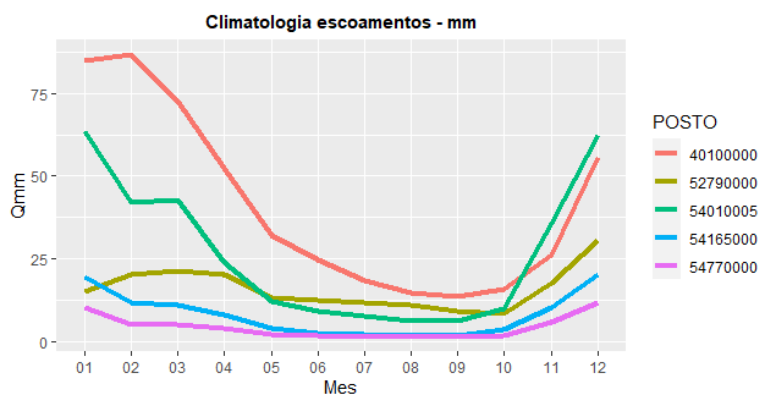
Fonte: Elaborado pelo autor

Figura 17-Climatologia de precipitações médias sobre as bacias selecionadas ($\text{mm}\cdot\text{mês}^{-1}$) elaborado a partir do produto CHIRPS período (1981-2020).



Fonte: Elaborado pelo autor

Figura 18-Climatologia de vazões ($\text{mm}\cdot\text{mês}^{-1}$) elaborado a partir dos dados disponibilizados pela ANA no período de 1980-2020



Fonte: Elaborado pelo autor

A variabilidade da vazão a ser prevista no mês de setembro, mês mais seco no Cerrado brasileiro, é apresentada na **Tabela 6** onde são apresentados os valores médio, mínimo, máximo e valores do primeiro e terceiro quartil da distribuição dos dados.

Tabela 6-Variabilidade das vazões (mm·mês⁻¹) no mês de setembro – valores mínimos, máximo, médio, primeiro quartil (1.Q) e terceiro quartil (3.Q) da amostra.

RIO	Mín.	Máx.	média	1.Q	3.Q
ALTO_SF	5,2	33,6	13,6	11,3	14,8
ALTO_JQ	2,2	13	6,3	4,2	8,7
VACARIA	0	7,5	1,8	0,4	7,5
Rib_SF	0	5,8	1,3	0,2	5,9
GONGOGI	0,9	22,8	9,2	4,4	10,3

Fonte: Elaborado pelo autor a partir de ANA (2022)

3.3 DADOS

3.3.1 Dados hidrológicos

Foram selecionados 5 postos fluviométricos cujos dados foram obtidos pelo sistema hidroweb do SNIRH (ANA, 2022). Dados classificados como “estimados” ou “duvidosos” não foram considerados. A despeito da disponibilidade de dados de cada posto, apresentada na **Tabela 7**, somente foram considerados dados a partir de 1981 até 2020 a fim de manter correspondência temporal com os dados de sensoriamento remoto do sensor AVHRR/NOAA. Não foram preenchidas falhas existentes, sendo esses meses desconsiderados da análise, exceto quando indicado no experimento. O período de dados efetivamente utilizado é apresentado na **Tabela 8**.

Tabela 7-Estações fluviométricas selecionadas para previsão.

RIO	Código	Início Operação	Período utilizado	Lat/Log	Área (km ²)	QMLT ⁽¹⁾ (mm/mês)
São Francisco	40100000	1956	1981-2018	-19,28/-45,28	14.000	41,5
Jequitinhonha	54010005	1980	1981-2019	-17,31/-43,21	7.720	26,8
Rio Vacaria	54165000	1975	1981-2017	-16,19/-42,59	2.290	8,0
Ribeirão São Francisco	54770000	1973	1981-2019	-16,12/-40,74	2.820	4,3
Rio Gongogi	52790000	1947	1981-2019	-14,33/-39,08	5.270	15,9

Nota: Média realizada no período utilizado; elaborado pelo autor a partir de ANA (2022).

Tabela 8-Disponibilidade de dados no mês de setembro (Q_{SET})

RIO	NOME	Início Operação	Período utilizado	Anos sem dados
São Francisco	ALTO_SF	1956	1981-2018	1989/1994
Jequitinhonha	ALTO_JQ	1980	1981-2019	1986/1992/1994
Rio Vacaria	VACARIA	1975	1981-2017	1994
Ribeirão São Francisco	Rib_SF	1973	1981-2019	1989/1999/1993/1994
Rio Gongogi	GONGOGI	1947	1981-2019	1993/1994

Nota: Elaborado pelo autor a partir de ANA (2022)

3.3.2 Dados de Precipitação

Os dados de precipitação utilizados foram adquiridos do produto *Climate hazards infrared precipitation with stations* - CHIRPS (FUNK et al., 2015) obtidos a partir do GEE (GORELICK et al., 2017). Foi escolhido por ser um produto de alta resolução $0,05^\circ$ e incorpora dados de climatologia globais, produtos vários satélites e dados medidos de precipitação em estações. Desta forma, o algoritmo CHIRPS abrange 3 grandes componentes: (1) Climatologia CHPclim obtida através de normais climatológicas, médias de satélites, dados de elevação e correlação espacial; (2) os dados do satélite *Climate Hazards Group Infrared Precipitation* (CHIRP) e (3) o procedimento de “*blending*” dos dados de satélite com os dados observados em estações, que inclui correção de viés e interpolação.

No Brasil o produto CHIRPS foi testado no Nordeste e apresentou uma correlação ($r=0,94$) em todas as estações testadas, porém com capacidade de detecção pequena nessa região, e tende a superestimar precipitações mensais maiores que $100\text{mm}\cdot\text{mês}^{-1}$ e subestimar menores. Seu melhor desempenho global foi no bioma Cerrado, sendo um produto elegível para substituição de precipitações fora do semiárido (PAREDES-TREJO; BARBOSA; LAKSHMI KUMAR, 2017).

3.3.3 Índice Palmer

O índice de seca Palmer foi utilizado na sua forma meteorológica (*Palmer Drought Severity Index – PDSI*). Foram obtidos a partir da base de dados *Terra Climate* (ABATZOGLOU et al., 2018) disponibilizados no GEE (GORELICK et al., 2017) em resolução temporal mensal, e em resolução espacial de $4,6\text{ km} \times 4,6\text{ km}$ por pixel.

Este índice desenvolvido por PALMER (1965) trata-se de um índice baseado em um balanço hídrico entre suprimento e retiradas de umidade em duas camadas de solo, sendo, portanto, uma medida padronizada do teor de umidade do solo (WEBER; NKEMDIRIM, 1998) e apresenta a vantagem de ser comparável espacialmente e temporalmente.

Este índice apresenta a característica de ser bastante persistente e auto correlacionado (ALLEY, 1984) porém em sua forma meteorológica pode antecipar 1 mês mudanças na umidade do solo e 90-120 dias os efeitos da seca hidrológica (WEBER; NKEMDIRIM, 1998).

O PDSI também tem mostrado fortes correlações com os deflúvios anuais, mesmo quando disponibilizado em escala global (DAI, 2011a). ZHAI et al. (2010) também encontraram correlações significativas deste índice com os coeficientes de escoamento (*run-off*) em escala anual em grandes bacias hidrográficas.

3.3.4 Índices VCI e TCI

Os índices VCI-TCI desenvolvidos por KOGAN (1995) são uma sólida fonte de informação sobre o stress hídrico da vegetação, a qual a seca é a maior causa. Ainda, pode ser considerada bastante útil como uma forma de diagnóstico em tempo real das condições da vegetação especialmente se dados meteorológicos de monitoramento não estiverem disponíveis.

Este índice também tem sido testado e empregado com sucesso no monitoramento dos impactos provocados pela seca na produção agrícola em todo o mundo (KOGAN, 1997), e mostrado também habilidade em antecipar eventos de seca (KOGAN; GUO, 2016), tanto em larga como pequena escala (KOGAN, 1995).

Os dados foram obtidos pela STAR – NOAA (https://www.star.nesdis.noaa.gov/pub/corp/scsb/wguo/data/Blended_VH_4km/VH/) e consiste num produto mesclado (Blended -VHP – *Vegetation Health Product*) derivado do sensor VIIRS (desde 2013) e do AVHRR (1981-2012). Os arquivos são disponibilizados no formato “netcdf”, com uma resolução espacial de 4km x 4km e disponibilizados semanalmente.

Os arquivos em resolução temporal semanal foram baixados e processados individualmente (extração e conversão para o formato tabular).

3.4 ANÁLISE DE TENDÊNCIAS

Esta etapa foi realizada com o objetivo de identificar mudanças de tendência nas séries de variáveis hidrológicas e climatológicas empregadas no experimento para fim de detectar possíveis períodos com tendências distintas entre os preditores e a vazão de setembro, bem como investigar suas causas.

Uma provável causa de modificação de tendências é a antropização da série histórica de vazões observadas em virtude de desmatamento, implantação e operação de reservatórios, transposições de água, e retiradas por usos consuntivos (ONS, 2007), e nestes casos podem induzir a um aprendizado distorcido dos modelos ML sobre dados impactados.

Para identificação de mudanças de tendências nas séries, foi aplicado o teste não paramétrico de Mann-Kendall (MK), largamente aplicado para detectar acréscimos ou decréscimos monótonos nos dados, que é aplicável mesmo diante da ocorrência de um evento abrupto ou episódico na série (MCCUEN, 2003). A hipótese nula (H_0) do teste é que os dados são oriundos de uma população com eventos independentes e uniformemente distribuídos, enquanto a hipótese alternativa (H_1) é que os dados seguem uma tendência monotônica. MK tem sido o teste mais aplicado e recomendado para avaliação de mudanças de tendência em séries temporais de dados climatológicos e hidrológicos (ROSIN; AMORIM; MORAIS, 2016;ZHANG et al., 2009)

Para identificar o ponto de mudança de tendência das variáveis foi utilizado o teste não-paramétrico de Pettitt descrito em BACK (2001) e MAINTAINER; POHLERT (2018). Diversas aplicações desse teste têm sido realizadas no sentido de estimar o ponto de mudança de tendência em series históricas de vazão e precipitação (Zhang, 2009; Back, 2001; Uliana et. al., 2014). A hipótese nula do teste de Pettit é que não há mudança na tendência central da série, com provável ponto de mudança K.

Os testes foram aplicados com o uso da biblioteca **trend** (Pohlert, 2018) do software R (RSTUDIO TEAM, 2022) em todas as séries de preditores. Nenhum tipo de redução de ruído prévio remoção de autocorrelação ou sazonalidade foi realizada.

Para melhor identificar as tendências nos dados foi realizada uma suavização de ruído nos dados com a regressão do tipo *loess* que ajusta uma função polinomial localizada. Esse procedimento foi realizado com a biblioteca e **ggplot2** (WICKHAM, 2016).

3.5 ANÁLISE DE CORRELAÇÕES

Preliminarmente ao ajuste dos modelos, foi realizada uma análise de correlações, no sentido de identificar previamente colinearidade e orientar a melhor seleção dos preditores.

Não obstante se tenha escolhido os preditores a partir de sua relevância física no processo hidrológico e os processos e modelos empregados sejam não-lineares, foi realizada uma análise de correlação linear cruzada entre todas as variáveis preditoras e alvo. Para isso se utilizou a técnica da matriz de correlação cruzada, com o emprego da biblioteca **coorplot** (WEIN; SIMKO, 2017) que fornece uma matriz gráfica contendo os valores de coeficiente de correlação, funções de densidades de distribuição dos dados e nível de significância do ajuste.

Uma vez que se espera que o PDSI possa desempenhar um papel substituto da vazão observada em bacias não monitoradas, foi realizada uma análise de correlação em múltiplas frequências utilizando-se a transformada de Wavelet, uma ferramenta apropriada para se obter *insights* e derivar relações empíricas entre os dados mesmo em processos não estacionários. Essa transformada decompõe uma série temporal no domínio tempo-frequência e identifica oscilações intermitentes com diferentes periodicidades (GRINSTED; MOORE; JEVREJEVA, 2004), o que não é possível simplesmente com a correlação serial de Pearson.

De acordo com MARAUN; KURTHS (2004), o tipo de transformação recomendada para analisar o grau de associação (coerência) entre dois sinais não lineares é a *Wavelet Transform Coherence* (WTC). Essa transformação foi aplicada com o uso da biblioteca **biwavelet** (GOUHIER; GRINSTED; SIMKO, 2019) com o software Rstudio.

A análise de Wavelet contínua *Continuous Wavelet Transform* - CWT também foi empregada sobre o sinal da vazão mínima Q_{SET} de forma subsidiária a análise dos experimentos. Essa transformação foi realizada diante da necessidade de compreensão do desempenho do modelo perante as características de cada sinal.

Também foram realizados estudos de autocorrelação com o emprego das funções ACF (autocorrelograma) e PACF (autocorrelograma parcial) a fim de subsidiar a seleção de melhores preditores com base no modelo autorregressivo integrado de médias móveis ARIMA (COGHLAN, 2016), para tal foi utilizada a biblioteca TTR (ULRICH, 2020).

3.6 MODELAGEM ML

3.6.1 Transformações de dados

Todas as variáveis foram adimensionalizadas pela sua média de longo período da variável, conforme **Equação 29** a **Equação 32** apresentadas a seguir. Essa transformação foi realizada com o objetivo de facilitar a análise do potencial de generalização do modelo, uma vez que admitindo-se que ela seja possível, a sua extensão em bacias sem dados pode ser cogitada através de técnicas de regionalização hidrológica.

Os índices VCI e TCI foram transformados (divididos) por 100 e o índice PDSI não sofreu nenhum tipo de transformação, exceto a executada no experimento Exp3 para obtenção da função empírica ortogonal EOF.

$$Q_{set}^*(t) = \frac{Q_{set}(t)}{Q_{set}} \quad \text{Equação 29}$$

$$P_{ond}^*(t) = \frac{P_{ond}(t)}{P_{ond}} \quad \text{Equação 30}$$

$$Q_{ond}^*(t) = \frac{Q_{ond}(t)}{Q P_{ond}} \quad \text{Equação 31}$$

$$Q_{ondjfm}^*(t) = \frac{Q_{ondjfm}(t)}{Q_{ondjfm}} \quad \text{Equação 32}$$

3.6.2 Algoritmos empregados

A modelagem envolveu o planejamento experimental, escolha dos algoritmos ML empregados na estimativa do modelo, calibração (*tunning*), escolha dos preditores mais relevantes, avaliação da acurácia, comparação entre os modelos, avaliação da capacidade de generalização dos modelos sobre dados novos e em bacias com características similares.

Foram escolhidos três algoritmos *machine learning*: Máquinas de Vetor de Suporte (*Support Vector Machine- SVM*), RF (*random forest*) e XGBoost (*Extreme*

Gradiente Boosting) apresentando em profundidade na **seção 2.4**, sendo que nos experimentos iniciais também foi utilizado o modelo linear (*lm*).

O algoritmo SVM foi desenvolvido por CORTES; VAPNIK (1995) sendo uma técnica fundamentada na teoria do aprendizado estatístico (VAPNIK, 1995) e otimiza o risco estrutural, tem sido amplamente aplicado no campo da hidrologia em diversas aplicações, e apresenta como característica o elevado poder de generalização, identificação de padrões e garantia de um ótimo global. Funcionando bem em amostra lineares e não-linearmente separáveis.

O algoritmo RF foi desenvolvido por BREIMAN (2001) e consiste na associação de árvores de decisão (*decision trees*) com uma poderosa técnica de reamostragem (*Bootstrap Agreggation*) que permite contornar o sobreajuste (*overfitting*) através de múltiplas visões e treino sobre subamostras de dados. Trata-se, portanto, de um conjunto de árvores de decisão treinadas paralelamente sobre inúmeras subamostras, fazendo com que se tenha um conjunto de árvores (*ensemble decision trees*). A decisão de todos os modelos é considerada na previsão final através da média, o que reduz a variância do modelo final formado pelo conjunto. Esse algoritmo tem seu maior campo de aplicação na classificação de imagens multiespectrais, entretanto também já foi aplicado em séries temporais no campo das ciências hidrológicas.

O algoritmo XGBoost (CHEN; GUESTRIN, 2016) é considerado o estado-da-arte dos algoritmos ML que não utilizam a técnica de aprendizado profundo (*Deep Learning*). Esse modelo utiliza a técnica de *Boosting* que consiste na adição sucessiva de modelos fracos (*weak learners*) visando corrigir os erros do conjunto de modelos adicionados até o passo anterior. O algoritmo XGBoost utiliza a otimização da função de perda com base na técnica do gradiente e minimiza o risco estrutural uma vez que na sua função de perda também há uma parcela de regularização.

3.6.3 Descrição dos experimentos

Para atingir os objetivos dessa pesquisa, foram realizados 7 experimentos apresentados na **Tabela 9** conforme explicado a seguir.

O experimento **Exp1** teve como objetivo investigar a correlação linear entre as variáveis preditoras P, Q, VCI, TCI e PDSI conhecidas ao término da estação chuvosa ONDJFM. Esse experimento gerou um modelo linear calibrado individualmente para cada bacia experimental, para todo o período de dados disponível e teve como objetivo realizar uma avaliação dos preditores quanto ao acréscimo de variância, medida pelo

índice VIF (*Variance Inflation Factor*) que estima a inflação da variância em modelos lineares e logísticos. No experimento Exp1 se considerou todo o período disponível de dados em cada bacia sem distinção entre amostra de teste e treino.

O VIF é um método complementar e mais robusto de identificação de colinearidade que a matriz de correlação, uma vez que permite a identificação de colinearidade múltipla entre os preditores (multicolinearidade) e foi calculado para cada preditor com o uso da biblioteca **car** (FOX; WEISBER, 2019), sendo uma medida de o quanto esse preditor pode ser calculado por uma combinação linear dos outros preditores. De modo geral valores de VIF maiores que o intervalo de 5-10 são considerados elevados (JAMES et al., 2013).

O VIF é calculado pela fórmula

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_{x_j|x_{-j}}^2} \quad \text{Equação 33}$$

Onde $R_{x_j|x_{-j}}^2$ é o coeficiente de regressão do preditor j em relação aos demais preditores, portanto o quanto mais próximo de 1 maior será o VIF, e consequentemente a colinearidade. Uma maneira de resolver o problema é a retirada das variáveis da base de preditores.

O experimento **Exp2** teve como objetivo comparar as performances individuais de cada algoritmo na atividade de previsão, considerando 6 e 9 meses de antecedência.

O experimento **Exp3** foi realizado com o objetivo de se verificar a influência da variabilidade espacial nos resultados. Os experimentos Exp1 e Exp2 consideraram preditores calculados como valores médios por bacia. Neste experimento foi considerada a variabilidade espaço temporal dos pixels inseridos em cada bacia hidrográfica. Para esse experimento foram extraídos os principais fatores utilizando a técnica de EOF's, escolhendo os fatores cuja soma explicasse aproximadamente 70% da variabilidade nos dados. Esse experimento foi aplicado na bacia de maior área de contribuição: ALTO_SF e foi executado para 6 meses de antecedência.

O experimento **Exp4** foi planejado após a avaliação de performance e análise dos resultados dos Exp1 a Exp3, quando o baixo desempenho verificado para algumas bacias demandou a análise mais aprofundada do sinal da vazão mínima de setembro $Q_{SET}^*(t)$, bem como a influência dos valores em intervalos anteriores. Para isso foram realizadas decomposições da variável $Q_{SET}^*(t)$ nos domínios do tempo e da frequência

através das transformações de wavelets (CWT) permitindo identificar regiões do espectro com maior potência do sinal.

Após a análise CWT foram analisadas as funções de autocorrelação ACF e PACF das séries temporais de Q_{SET}^* e PDSI, cujos resultados são apresentados na **Figura 38**. A partir das análises das funções ACF e PACF, foram acrescentados os seguintes preditores:

- sinal suavizado da média móvel de 3 anos da vazão mínima Q_{SET}^* ($Q_{SET}.SMA.3$);
- Vazão de Q_{SET}^* no ano anterior ($Q_{SET}^*_{t-1}$), em algumas bacias de forma experimental foi acrescentada a vazão $Q_{SET}^*_{t-2}$, como foi o caso da bacia do rio Rib_SF e $Q_{SET}^*_{t-3}$ no rio GONGOI;
- Vazão do período chuvoso do ano anterior Q_{ONDJFM_lag1} ;
- A derivada de 1ª ordem do sinal suavizado ($Q_{SET}^*.SMA.3.diff$) para representar a componente de mudança de tendência do sinal e remover a ausência de não-estacionariedade, prejudicial para o desempenho dos algoritmos ML (TAN; PERKOWSKI, 2015). Após a diferenciação da série de Q_{SET}^* , o teste estatístico de Dick-Fuller foi executado com a biblioteca **urca** (PFAFF, 2008) como qual se conferiu estacionariedade em todas as séries diferenciadas. Testes preliminares mostraram que a diferenciação sobre o sinal suavizado contribui com a obtenção de modelos mais acurados do que sobre o sinal original de Q_{SET}^* , sendo esse o sinal adotado durante o treinamento.

O experimento **Exp5** foi realizado com o objetivo de verificar a habilidade do índice PDSI ser um substituto da vazão observada em bacias sem monitoramento e fez uso das funções ACF e PAC par melhor estimar correlações temporais deste indicador.

O experimento **Exp6** foi realizado para testar a generalização do modelo ML que apresentou melhor desempenho, em bacias vizinhas com disponibilidade de dados hidrológicos. Neste experimento apenas o modelo foi transferido, utilizando-se os preditores da bacia local.

O experimento **Exp7** foi realizado de maneira similar ao **Exp6**, entretanto além da transferência do modelo ML, também são transferidas todas as variáveis adimensionais relacionadas à vazão, mantendo-se os preditores locais derivados por sensoriamento remoto. Este experimento teve o objetivo de testar a generalização em bacias sem disponibilidade de dados hidrológicos.

Tabela 9-Descrição dos experimentos realizados em que os modelos são ajustados tendo como entrada as variáveis preditoras e saída a vazão média no mês de setembro (Q_{SET}^*)

Experim.	Objetivo	Anteced.	Preditores	Modelo	Bacias
Exp1	Investigação do grau de correlação linear entre as variáveis e da ocorrência de multicolineariedade	6 meses	P_{ONDJFM}^* , Q_{ONDJFM}^* , VCI_{ONDJFM}^* , TCI_{ONDJFM}^* , $PDSI_{ONDJFM}$	lm (linear model)	TODAS Individualmente
Exp2	Comparação entre a habilidade de algoritmos ML	6 e 9 meses	P_{OND}^* , Q_{OND}^* , VCI_{OND}^* , TCI_{OND}^* , $PDSI_{OND}$, P_{ONDJFM}^* , Q_{ONDJFM}^* , VCI_{ONDJFM} , TCI_{ONDJFM} , $PDSI_{ONDJFM}$	lm SVM RF XGBoost	TODAS Individualmente
Exp3	Investigação do ganho de acurácia pela consideração da variabilidade espacial através das extrações das funções empíricas ortogonais (EOFs)	6 meses	EOF's obtidas pela rotação varimax, sobre os dados em pixel: P_{ONDJFM}^* , VCI_{ONDJFM} , TCI_{ONDJFM} , $PDSI_{ONDJFM}$	lm SVM RF XGBoost	ALTO_SF
Exp4	Investigar a melhoria de desempenho considerando indicadores sugeridos por processos autorregressivos (ARIMA) e CWT	6 meses	Q_{ONDJFM}^* , P_{ONDJFM}^* , VCI_{ONDJFM} , TCI_{ONDJFM} , $PDSI_{ONDJFM}$, $Q_{ONDJFM\ t-1}^*$, $Q_{SET\ t-1}^*$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$	SVM RF XGBoost	TODAS Individualmente
Exp5	Retirada da vazão Q da base de preditores – Avaliar a performance do Palmer-PDSI como proxy, considerando indicadores sugeridos por processos autorregressivos (ARIMA) e CWT	6 meses	P_{ONDJFM} , VCI_{ONDJFM} , TCI_{ONDJFM} , $PDSI_{ONDJFM\ t, t-1, t-2, t-3}$, $PDSI_{ONDJFM.SMA.3}$, $PDSI_{ONDJFM.diff}$	SVM RF XGBoost	TODAS Individualmente
Exp6	Testar o poder de generalização dos modelos através de testes de validação cruzada em bacias distintas com dados locais	6 meses	Q_{ONDJFM}^* , P_{ONDJFM}^* , VCI_{ONDJFM} , TCI_{ONDJFM} , $PDSI_{ONDJFM}$, $Q_{ONDJFM\ t-1}^*$, $Q_{SET\ t-1}^*$, $Q_{SET\ t-2}^*$, $Q_{SET\ t-3}^*$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$	SVM	$\frac{\text{Alto_JQ} \otimes \text{ALTO_SF}}{\text{Alto_JQ} \otimes \text{Gongogi}}$ $\frac{\text{Vacaria} \otimes \text{Rib_SF}}$
Exp7	Testar o poder de generalização dos modelos através de testes de validação cruzada em bacias distintas sem dados locais de vazão	6 meses	Q_{ONDJFM}^* , P_{ONDJFM}^* , VCI_{ONDJFM} , TCI_{ONDJFM} , $PDSI_{ONDJFM}$, $Q_{ONDJFM\ t-1}^*$, $Q_{SET\ t-1}^*$, $Q_{SET\ t-2}^*$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$	SVM	$\frac{\text{Alto_JQ} \otimes \text{ALTO_SF}}{\text{Alto_JQ} \otimes \text{Gongogi}}$ $\frac{\text{Vacaria} \otimes \text{Rib_SF}}$

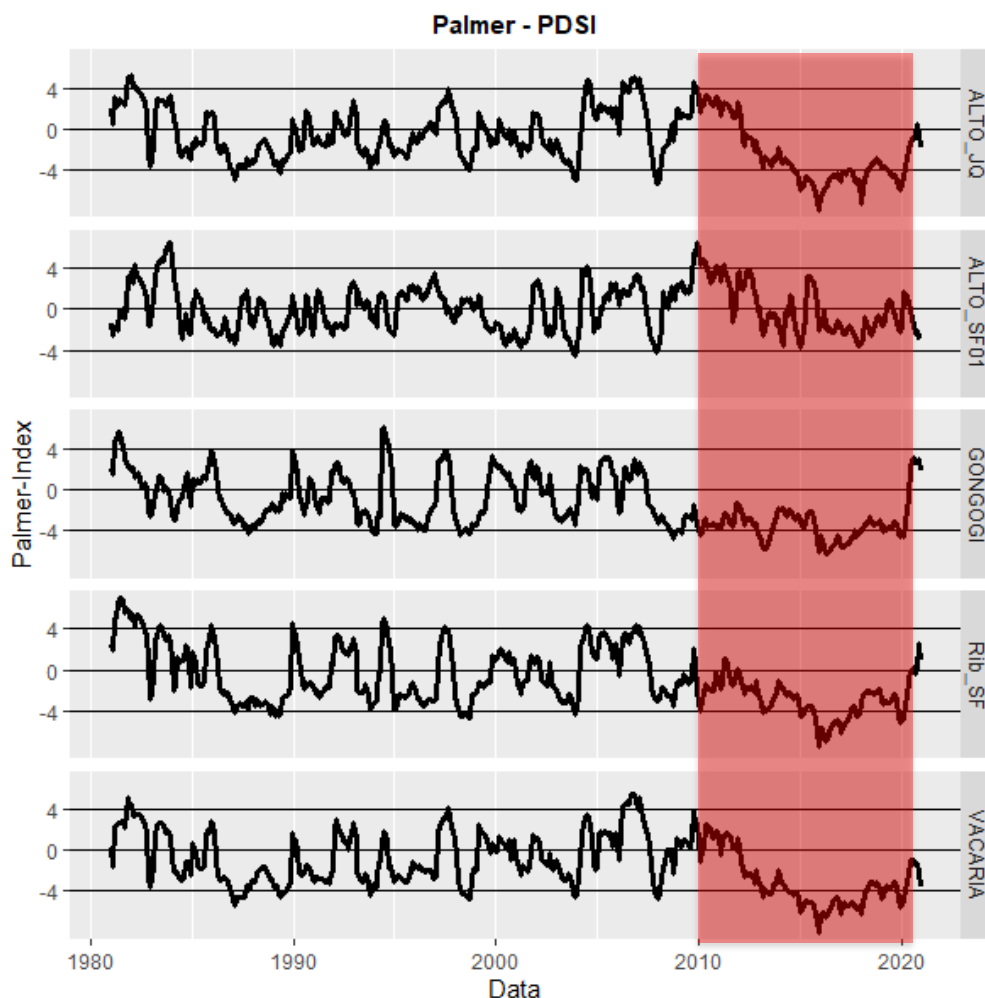
Nota: No experimento Exp4, considerou-se como preditor além da $Q_{SET\ t-1}^*$, a $Q_{SET\ t-2}^*$ na bacia Rib_SF e $Q_{SET\ t-3}^*$ na bacia do GONGOGI.

3.6.4 Separação da amostra de treino e validação

O período de treinamento do modelo em todos os experimentos foi escolhido baseando-se na análise do índice Palmer (PDSI), buscando-se excluir da amostra de treino o período da última seca. Analisando-se o gráfico do índice Palmer (**Figura 19**), observa-se que o índice atinge o nível “muito seco” ($PDSI < -3$) e “extremamente seco” ($PDSI < -4$) em tempos distinto em cada bacia. No rio Gongogi o índice atinge patamares de “muito seco” já no ano hidrológico 2008/2009 e na bacia do ALTO_SF no ano de 2013/2014.

Sendo assim, os experimentos buscaram realizar o treinamento para o período 1981-2010 deixando o período restante para validação. Alguma variação no período teste ocorreu em função da presença de falhas nos registros hidrológicos.

Figura 19-Variabilidade temporal índice Palmer (PDSI) em valores médios mensais por bacia no período de 1981-2020. Destaque em vermelho para o período de validação.



Fonte: Elaborado pelo autor a partir de ABATZOGLOU et al. (2018)

3.6.5 Calibração

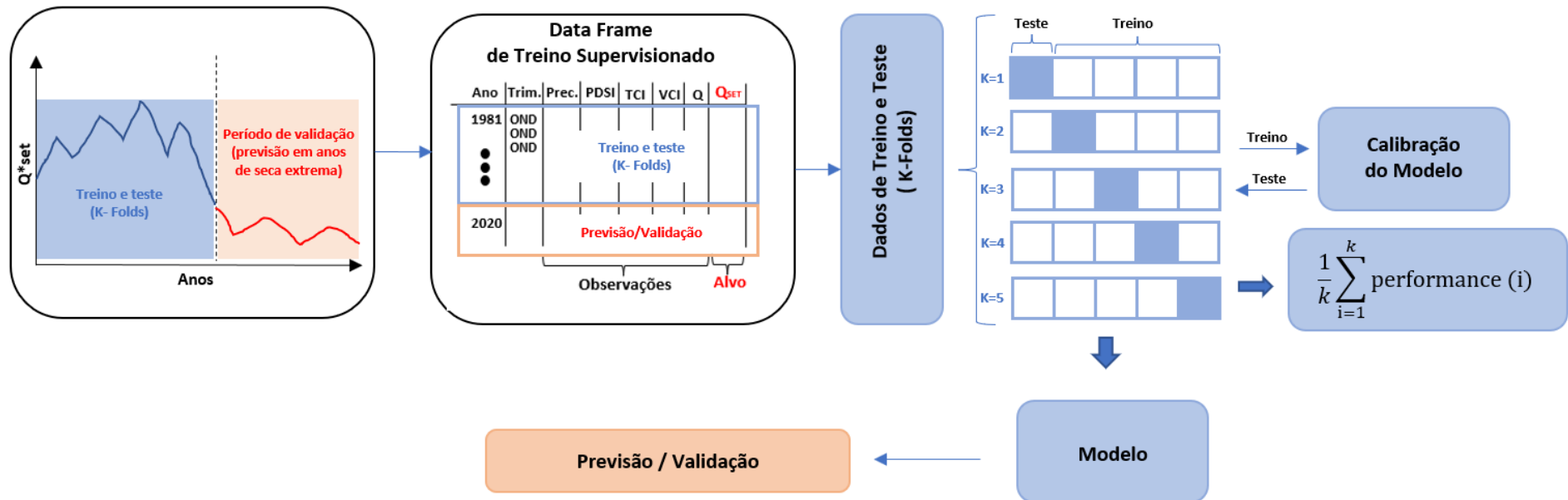
Os modelos foram calibrados utilizando o método de validação cruzada K-folds, que avalia a performance do modelo em diferentes subconjuntos da amostra de treinamento e calcula a média da acurácia na previsão. O algoritmo funciona da seguinte forma (KASSAMBARA, 2017):

- a amostra de treino é repartida de forma randômica em K-subconjuntos;
- retira-se um dos K subconjuntos e se realiza o treino nos demais K-1 subconjuntos;
- O algoritmo testa o modelo ajustado na amostra que foi retirada e guarda o valor da acurácia medida pelo RMSE (*root mean square error*);
- Repete esse processo até que todos os K subconjuntos tenham sido servidos como amostra de teste;
- O algoritmo computa a média dos erros calculados sobre os K subconjuntos, que é tomada como a métrica de performance do modelo.

Em todos os experimentos, se adotou K=5 (amostra dividida em 5 partes), e esse procedimento foi realizado de forma automática com o uso da biblioteca CARET (KUHN, 2020).

O treino envolveu os 18 primeiros anos da série, com o objetivo de não incluir o período da última seca na fase de treinamento, uma vez que um dos objetivos é avaliar como o modelo se comporta realizando previsões em um evento sem precedentes, e, portanto, não utilizado no treino. Devido à existência de falhas nos dados, ocorreu variação no tamanho da amostra de treino. O fluxo apresentado na **Figura 20** ilustra a metodologia descrita acima. A performance do modelo foi medida pelo RMSE calculado entre os dados previstos e observados.

Figura 20-Metodologia genérica de aplicação do método K-folds (k=5) na calibração dos modelos estatísticos.



Fonte: Elaborado pelo autor

4 RESULTADOS

Neste capítulo são apresentadas as análises de mudança de tendências e de correlação: linear, wavelets, ACF e PACF importantes para a caracterização prévia e subsídios da seleção de variáveis preditoras. Em seguida são apresentados em seções distintas os resultados experimentos Exp1 a Exp7 descritos no capítulo anterior.

4.1.1 *Análise de Estacionariedade e de Mudança de Tendências*

Os resultados dos testes de Man-Kendall e Pettitt são apresentados na **Figura 21**, considerando o período de 1981-2020 das variáveis precipitação (coluna 1), vazão média no período chuvoso ONDJFM (coluna 2), vazão mínima em setembro (coluna 3), índice Palmer PDSI (coluna 4), índice de temperatura TCI (coluna 5) e índice de vegetação VCI (coluna 6). Nas linhas (de cima para baixo) são as bacias: ALTO_SF (linha 1), ALTO_JQ (linha 2), VACARIA (linha 3), Rib_SF (linha 4) e GONGOGI (linha 5).

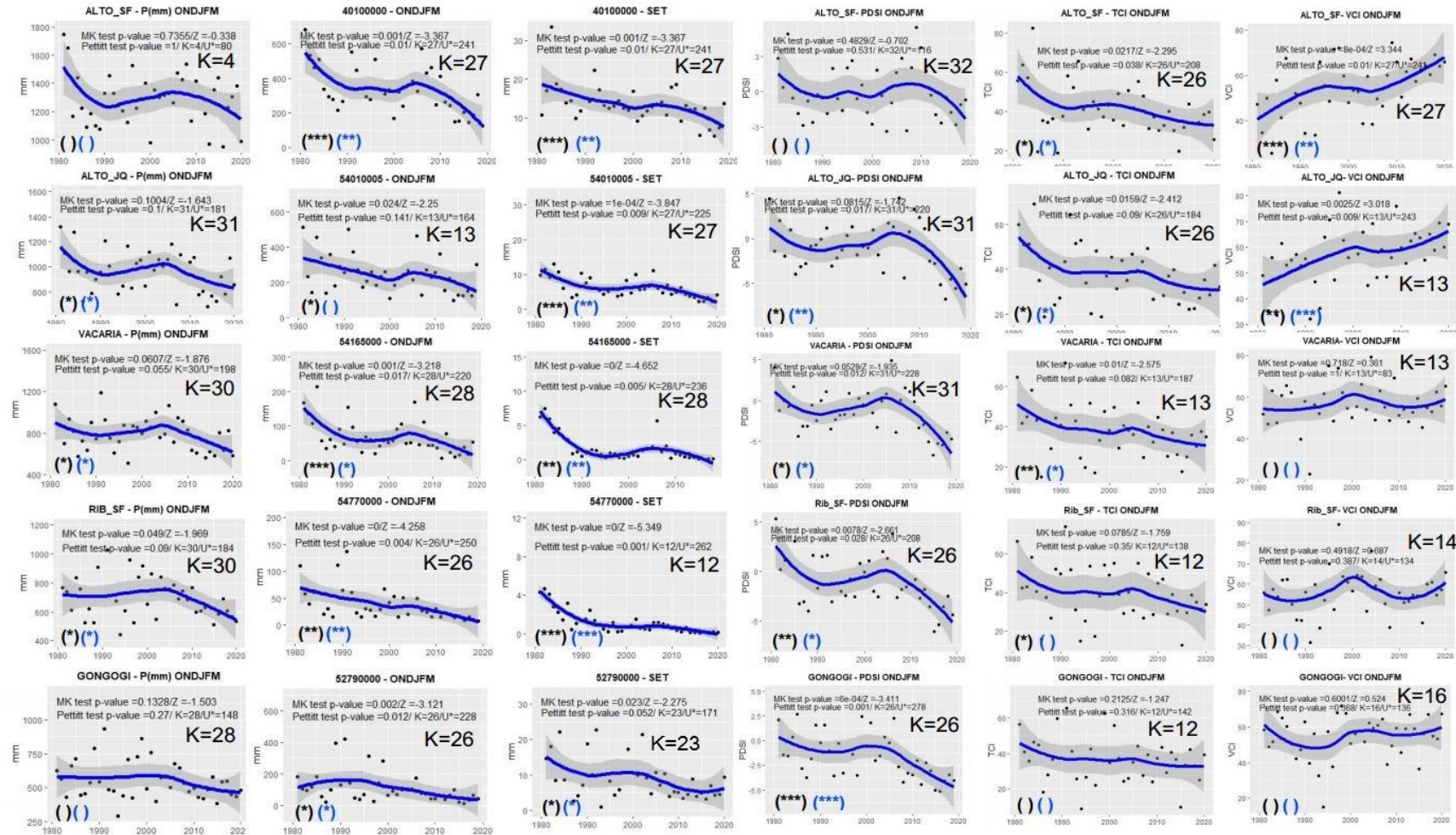
De modo geral, pode-se observar que todas as bacias apresentaram mudança de tendência significativa na vazão mínima (Q_{SET}) e essa mudança ocorreu na maioria dos casos entre os anos hidrológicos de 2007 e 2008. De forma análoga foi verificado para a vazão no período chuvoso Q_{ONDJFM} .

O sinal de precipitação P_{ONDJFM} também apresentou mudança de tendência significativa, exceto para as bacias ALTO_SF e GONGOGI. Tal mudança de tendência se apresenta com ponto provável de mudança no ano de 2010.

O sinal de TCI também apresentou mudanças significativas em todas as bacias exceto no GONGOGI. Os pontos de mudança deste sinal são anteriores às séries de precipitação e vazão, se situando ainda na década de 90 para as bacias dos rios VACARIA e Rib_SF. Ao se analisar conjuntamente os sinais de TCI e VCI tais mudanças parecem estar associadas a modificações no sinal da vegetação.

O índice Palmer (PDSI) que é representativo do balanço climatológico apresentou mudança de tendência significativa, com diminuição em todas as bacias, a partir de 2006 para as bacias Rib_SF e Gongogi localizadas na mata atlântica e 2011 em todo o Cerrado, exceto no ALTO_SF.

Figura 21-Tendências no período de 1981-2020 das variáveis precipitação (coluna 1), vazão média no período chuvoso ONDJFM (coluna 2), vazão mínima em setembro (coluna 3), índice Palmer PDSI (coluna 4), índice de temperatura TCI (coluna 5) e índice de vegetação VCI (coluna 6). Dados por bacia: entrada em linhas.



Legenda: Em preto a significância do teste Man-Kendall, em azul teste Pettitt. Os caracteres (*), (**), (***) representam os níveis de significância $p < 0,1$, $p < 0,01$, $p < 0,001$ respectivamente. Fonte: Elaborado pelo autor.

4.1.2 Análise de correlação

A análise de correlação cruzada entre os preditores e entre a variável alvo (Q_{SET}) e os preditores é apresentada na **Figura 22**. Esta análise considerou os dados de todas as bacias conjuntamente.

Nesta figura é possível observar que o grau de associação linear das variáveis preditoras e Q_{SET} são todas significativas ($p < 0,01$ e $p < 0,001$) sendo a correlação com a vazão do período chuvoso Q_{ONDJFM} a que se destaca com maior grau de associação, em segundo lugar vem a associação com a precipitação no período chuvoso P_{ONDJFM} .

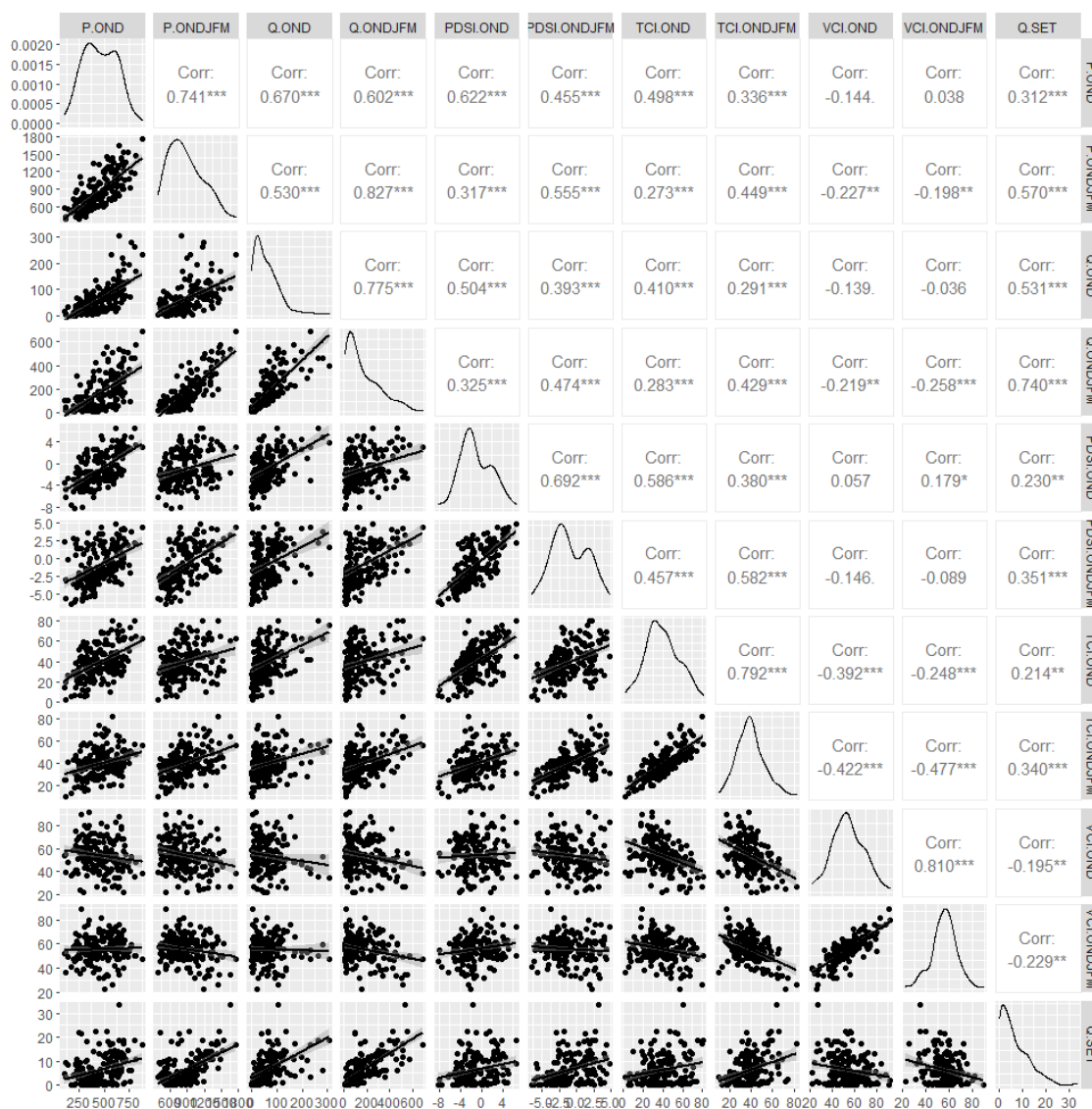
A relação de Q_{SET} com os índices PDSI e TCI também são positivas e significativas sendo a associação com o primeiro, levemente superior. Também se observa uma relação bastante forte de PDSI com a precipitação e com a vazão, a precipitação e Temperatura (TCI). O índice VCI se apresentou negativamente correlacionado com a vazão de Q_{SET} e com a precipitação.

A análise de correlação entre os sinais de intervalo mensal ($Q \times PDSI$) são apresentados na **Figura 23**. Analisando-se a figura é possível observar que existe uma correlação bastante significativa entre os dois índices na escala intra-anual (menor que 12 meses) em todas as bacias, entretanto essa correlação não é contínua no tempo. Observa-se também que existe uma correlação bastante significativa na escala de 32 meses em quase todas as bacias, exceto na bacia ALTO_SF. Observa-se também na **Figura 23** que as variáveis ($Q \times PDSI$) estão em fase e que o PDSI atua como forçante da variável Q .

Na **Figura 24** é apresentada a decomposição do sinal de Q_{SET} pela transformação CWT onde se observa que o sinal de vazões mínimas apresenta uma variabilidade bastante descontínua nos domínios do tempo e da frequência, sendo significativa em períodos isolados e distintos. Exceto na bacia do ALTO_SF todas as bacias apresentaram baixa variabilidade do sinal de vazões mínimas após o período de 2010. Observa-se também, uma variabilidade mais forte, entretanto não significativa na escala de 2-4 anos em todas as bacias.

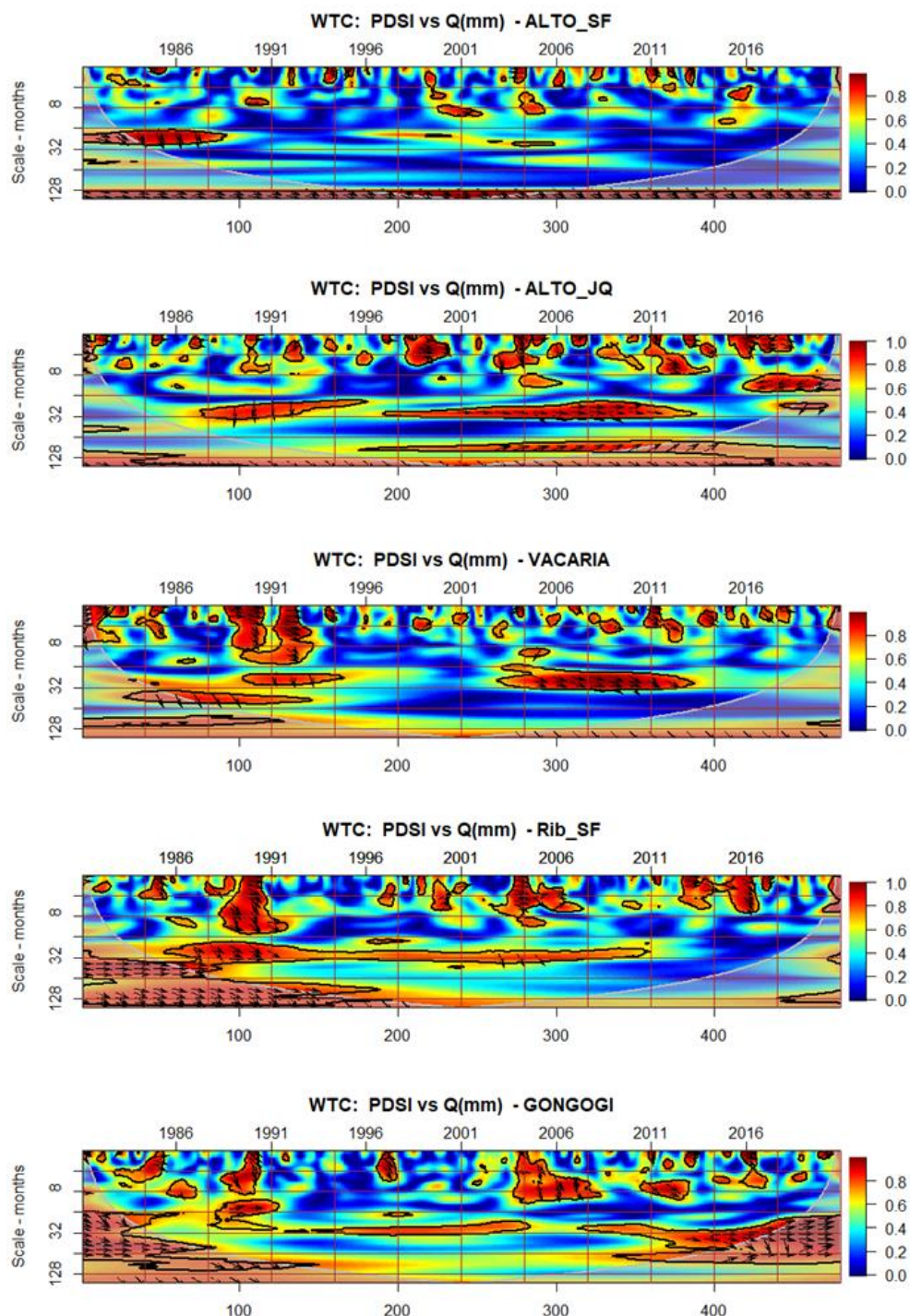
Na **Figura 25** são apresentados os gráficos das séries de variáveis adimensionais Q_{SET}^* , P_{OND}^* , P_{ONDJFM}^* , Q_{OND}^* , Q_{ONDJFM}^* , $PDSI_{OND}$, VCI_{OND} , TCI_{OND} onde se observa uma similaridade no padrão de variação destas variáveis na área de estudo. O teste de Tukey HSD foi aplicado (apêndice), podendo-se não rejeitar a hipótese de que as séries adimensionais pertencem à mesma população.

Figura 22-Matriz de correlações cruzadas entre as variáveis: P_{OND}, P_{ONDJFM}, Q_{OND}, Q_{ONDJFM}, PDSI_{OND}, PDSI_{ONDJFM}, TCI_{OND}, TCI_{ONDJFM}, VCI_{OND}, VCI_{ONDJFM}, Q_{SET}



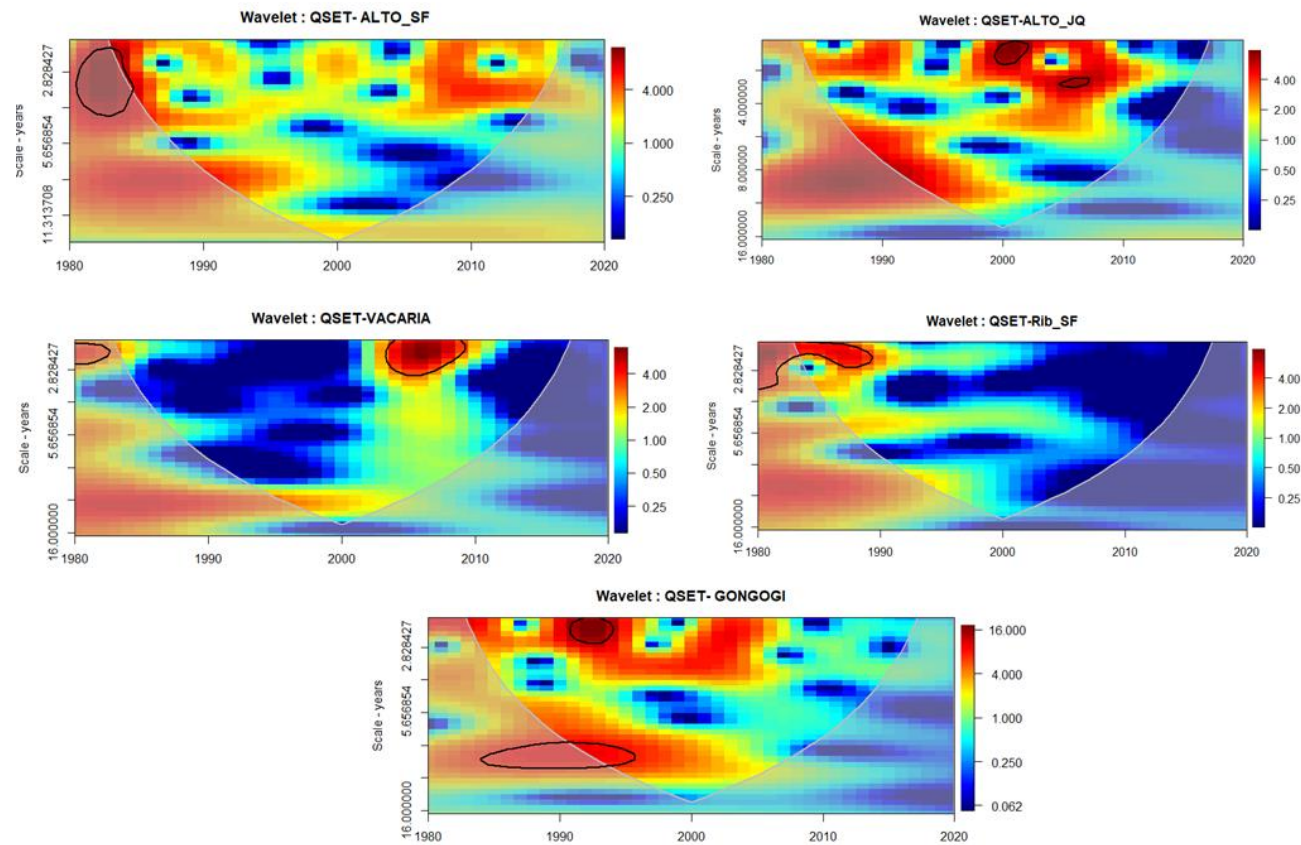
Nota: caracteres “*”, “**”, “***” representam o nível de significância da correlação linear $p < 0.1$, $p < 0.01$ e $p < 0.001$, respectivamente. Fonte: Elaborado pelo autor.

Figura 23-Transformações Wavelets-Coherence (WTC) entre os sinais do índice Palmer (PDSI) e as vazões observadas.



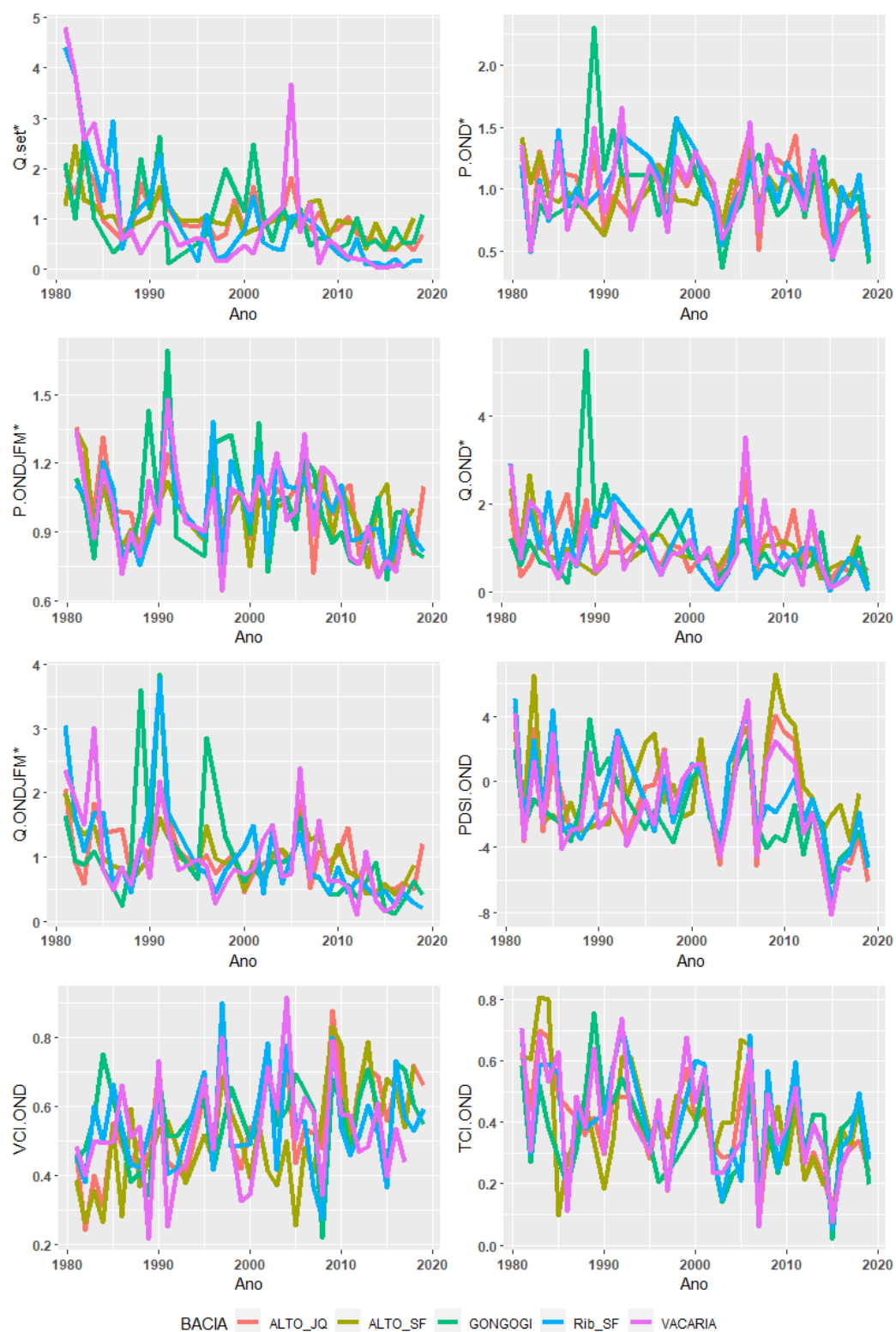
Nota: Escala em meses. Cores na paleta referem-se ao grau de associação entre os sinais mensais de PDSI e Q. Linhas sólidas pretas delimitam as regiões entre a correlação entre os sinais é significativa. Linha sólida branca delimita o cone de influência. Setas apontando para a direita significa que as series se movem em fase, para a esquerda em anti-fase. Setas apontando para a direita e para baixo indicam que PDSI está antecipando o sinal de Q. Fonte: Elaborado pelo autor.

Figura 24-Transformações Continuous Wavelet Transform (CWT) para a série de vazões mínimas Q_{SET}^* no mês de setembro, escala temporal em anos.



Nota: Escala em meses. Cores na paleta referem-se à potência do sinal. Linhas sólidas pretas delimitam as regiões onde a potência do sinal é significativa. Linha sólida branca delimita o cone de influência. Fonte: Elaborado pelo autor.

Figura 25-Variáveis adimensionais Q_{SET}^* , P_{OND}^* , P_{ONDJFM}^* , Q_{OND}^* , Q_{ONDJFM}^* , $PDSI_{OND}$, VCI_{OND} , TCI_{OND}



Fonte: Elaborado pelo autor

4.1.3 Resultados dos experimentos

4.1.3.1 Experimento Exp1 – análise de colinearidade de preditores

Os resultados do cálculo do fator de inflação da variância são apresentados na **Tabela 10** demonstrando que o grau de colinearidade entre os preditores é baixo uma vez que a maioria se manteve abaixo de 5 exceto para a precipitação na bacia ALTO_JQ, porém não ultrapassou o valor 10, de modo que nenhum tratamento adicional ou descarte de preditores foi realizado.

Tabela 10- Resultados experimento Exp1 –fator de inflação da variância (VIF) avaliada pelo modelo de regressão linear (lm) antecipando Q_{SET}^* com 6 meses de antecedência

BACIA	VIF				
	$PDSI_{ONDJFM}^*$	TCI_{ONDJFM}^*	VCI_{ONDJFM}^*	P_{ONDJFM}^*	Q_{ONDJFM}^*
ALTO_SF	1,72	2,77	2,36	2,70	3,52
ALTO_JQ	2,76	7,33	2,84	8,80	2,83
VACARIA	3,15	4,00	1,61	5,06	2,68
Rib_SF	3,37	2,80	1,26	3,03	2,75
GONGOGI	2,27	2,27	1,28	3,91	4,15

4.1.3.2 Experimento Exp2 – Análise de desempenho dos modelos utilizando preditores sazonais com antecedência de 6 e 9 meses

No experimento Exp2 foram empregados apenas preditores calculados a partir da estação chuvosa OND e ONDJFM dentro do mesmo ano hidrológico, sem considerar efeitos decorrentes de anos anteriores.

Os valores de performance RMSE e viés obtidos com a previsão dos modelos lm, SVM, RF e XGBoost são apresentados nas **Tabela 11** e **Tabela 12** para 9 e 6 meses de antecedência, respectivamente. E os gráficos referentes às previsões de Q_{SET}^* nas **Figura 26** a **Figura 36**.

Nas previsões com antecedência de 9 meses, o melhor desempenho foi conseguido com a aplicação do algoritmo SVM-Linear na bacia do ALTO_SF (**Figura 27**). Nas previsões de 6 meses o melhor desempenho foi conseguido pelo algoritmo XGboost na bacia do ALTO_JQ (**Figura 32**).

De forma geral, a performance do algoritmo SVM foi o algoritmo que derivou modelos com melhor performance.

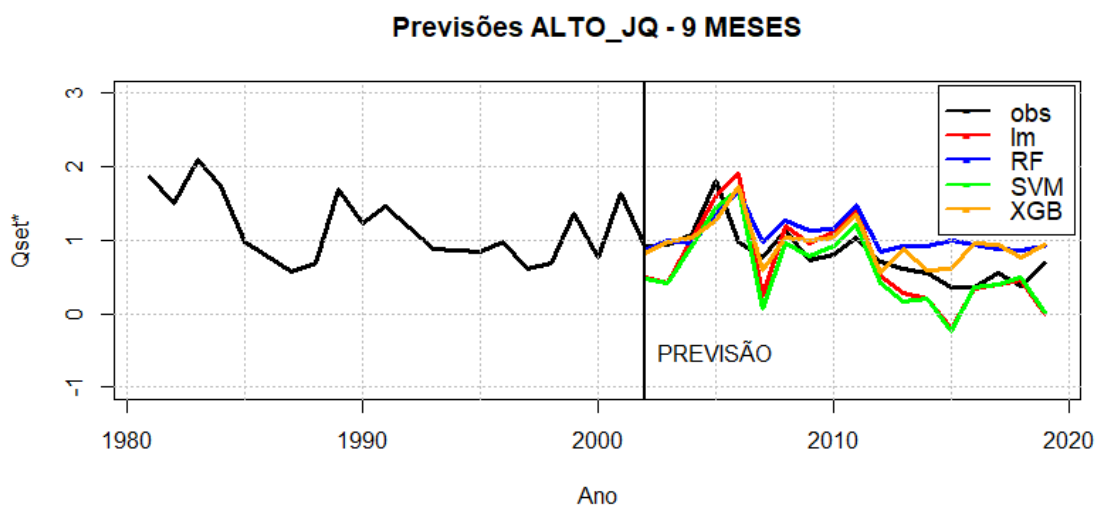
As piores performances aconteceram nas previsões dos rios VACARIA e no Rib_SF, e as melhores nos rios ALTO_SF e ALTO_JQ.

Tabela 11-Resultados experimento Exp2 – Acurácia nas amostras de teste e treino, antecedência de 9 meses na previsão de Q_{SET}^* com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM), XGBoost (XGB). Preditores utilizados: P_{OND}^* , Q_{OND}^* , $PDSI_{OND}$, VCI_{OND} , TCI_{OND}

BACIA	Treino	Teste	lm	SVM	RF	XGBoost
ALTO_SF	1981-2000	2001-2018	0,345 (0,363) [+0,185]	0,374 (0,251) [-0,080] Linear C=1	0,226 (0,399) [+0,289]	0,000 (0,420) [+0,230]
ALTO_JQ	1981-2001	2002-2019	0,263 (0,410) [-0,116]	0,286 (0,405) [-0,206] Linear C=1	0,208 (0,394) [+0,280]	0,000 (0,336) [+0,150]
VACARIA	1981-2001	2002-2017	0,883 (1,206) [-0,472]	1,02 (1,196) [+0,497] Linear C=1	0,645 (1,361) [-0,094]	0,000 (1,444) [+0,707]
Rib_SF	1981-2002	2003-2019	0,625 (1,557) [+1,110]	0,906 (0,882) [+0,793] Radial C=0,5/ Sigma=0,28	0,549 (1,326) [+1,176]	0,171 1,750 [+1,493]
GONGOGI	1981-2003	2004-2019	0,626 (0,601) [+0,287]	0,708 (0,488) [+0,023] Linear C=1	0,361 (0,499) [+0,375]	0,00 (0,722) [+0,337]

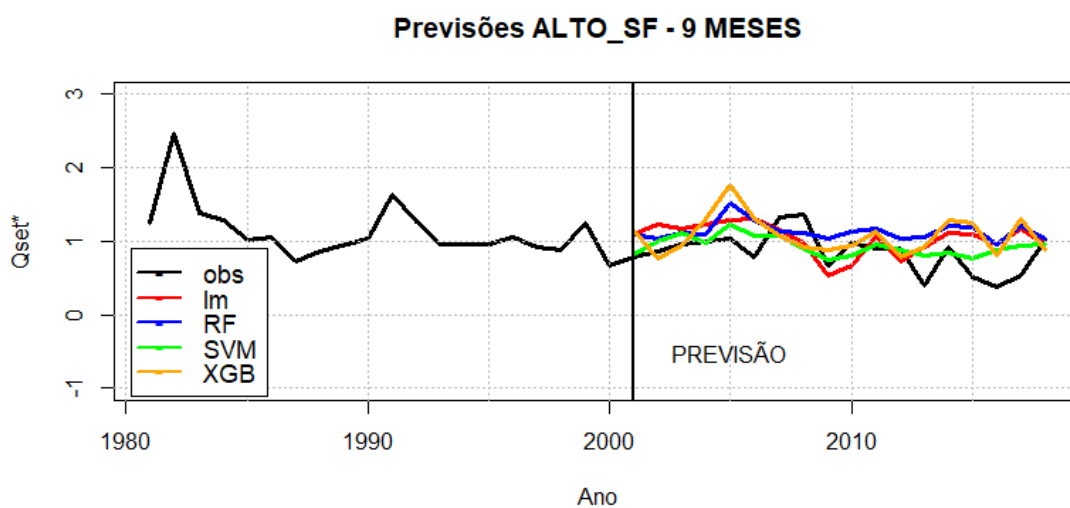
Nota: Entre parêntesis a acurácia (RMSE) apresentada na amostra de teste, fora do parêntesis a acurácia na fase de treino, entre colchetes o viés calculado na amostra de teste

Figura 26-Previsão de vazão Exp2 (Q_{SET}^*) com 9 meses de antecedência bacia ALTO_JQ com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{OND}^* , Q_{OND}^* , $PDSI_{OND}$, VCI_{OND} , TCI_{OND}



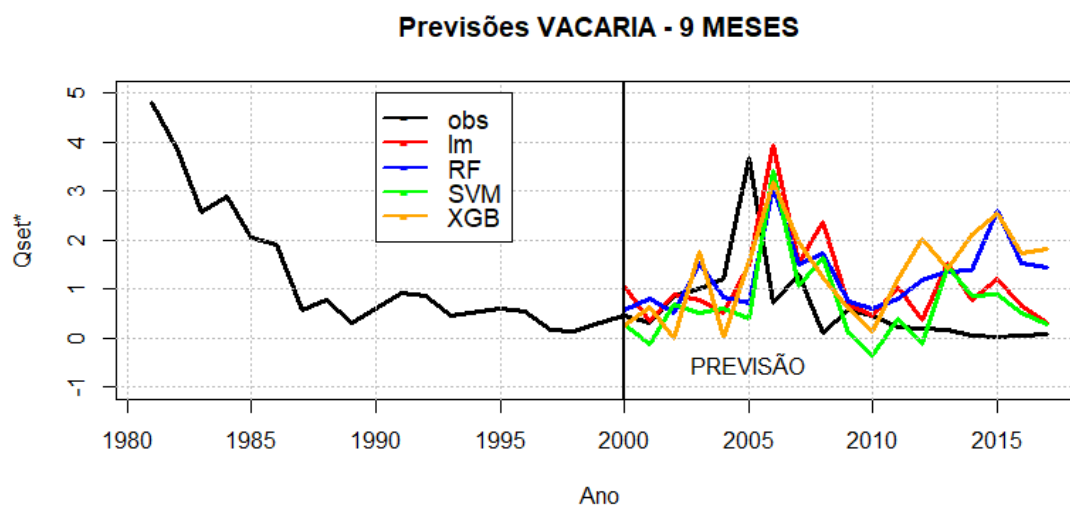
Fonte: Elaborado pelo autor.

Figura 27-Previsão de vazão Exp2 (Q_{SET}^*) com 9 meses de antecedência bacia ALTO_SF com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{OND}^* , Q_{OND}^* , $PDSI_{OND}$, VCI_{OND} , TCI_{OND}



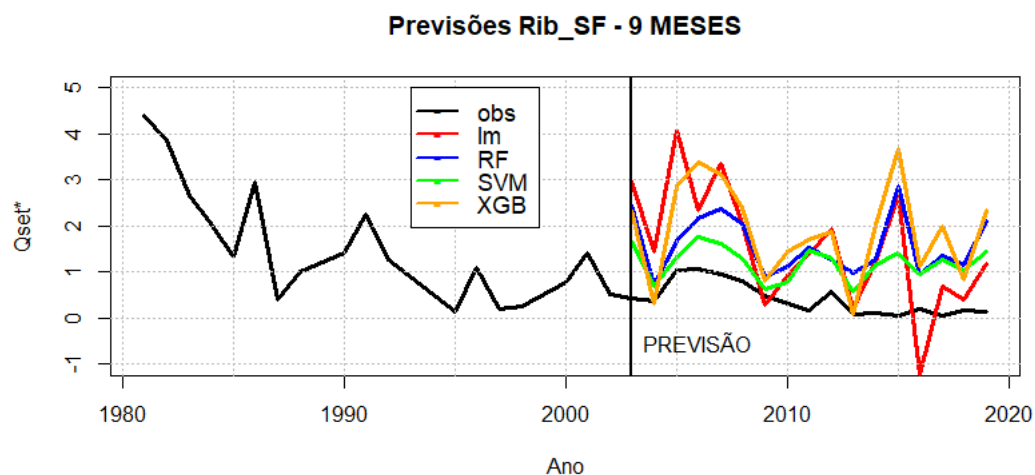
Fonte: Elaborado pelo autor.

Figura 28-Previsão de vazão Exp2 (Q_{SET}^*) com 9 meses de antecedência bacia VACARIA com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{OND}^* , Q_{OND}^* , $PDSI_{OND}$, VCI_{OND} , TCI_{OND}



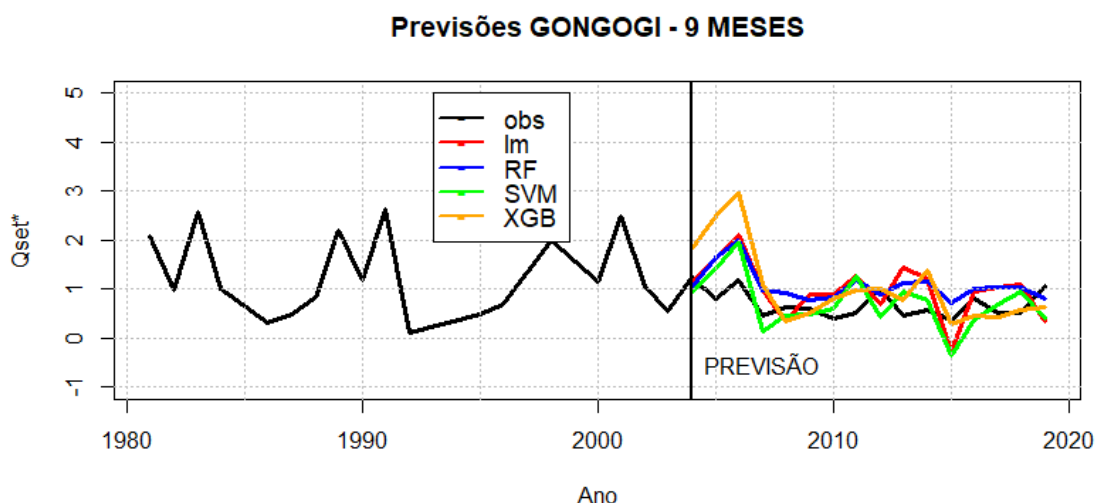
Fonte: Elaborado pelo autor.

Figura 29-Previsão de vazão Exp2 (Q_{SET}^*) com 9 meses de antecedência bacia Rib_SF com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM Radial), XGBoost (XGB). Preditores utilizados: P_{OND}^* , Q_{OND}^* , $PDSI_{OND}$, VCI_{OND} , TCI_{OND}



Fonte: Elaborado pelo autor.

Figura 30-Previsão de vazão Exp2 (Q_{SET}^*) com 9 meses de antecedência bacia GONGOGI com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{OND}^* , Q_{OND}^* , $PDSI_{OND}$, VCI_{OND} , TCI_{OND}



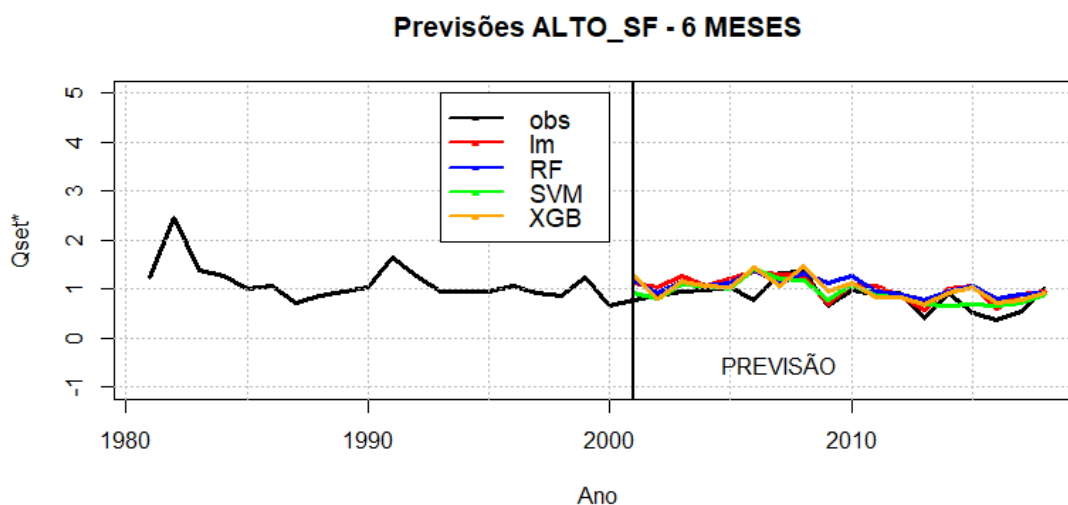
Fonte: Elaborado pelo autor.

Tabela 12-Resultados experimento Exp2 – Acurácia nas amostras de teste e treino, antecedência de 6 meses na previsão de Q_{SET}^* com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{ONDJFM}^* , Q_{ONDJFM}^* , $PDSI_{ONDJFM}$, VCI_{ONDJFM} , TCI_{ONDJFM}

BACIA	Treino	Teste	lm	SVM	RF	XGBoost
ALTO_SF	1981-2000	2001-2018	0,270 (0,262) [+0,178]	0,302 (0,216) [+0,07] Linear C=1	0,204 (0,306) [+0,203]	0,000 (0,282) [+0,159]
ALTO_JQ	1981-2001	2002-2019	0,346 (0,291) [+0,156]	0,363 (0,323) [+0,207] Linear C=1	0,194 (0,290) [+0,171]	0,000 (0,213) [+0,026]
VACARIA	1981-2001	2002-2017	0,851 (1,136) [0,000]	1,025 0,955 [-0,167] Linear	0,634 (1,207) [+0,426]	0,000 (1,493) [+0,716]
Rib_SF	1981-2002	2003-2019	0,700 (1,226) [+0,930]	1,065 (0,705) [+0,641] Radial C=0,25 Sigma=0,45	0,535 (1,085) [+0,908]	0,012 (1,06) [+0,729]
GONGOGI	1981-2003	2004-2019	0,538 (0,679) [+0,370]	0,565 (0,650) [+0,263] Linear C=1	0,334 (0,3380) [+0,202]	0,000 (0,450) [+0,102]

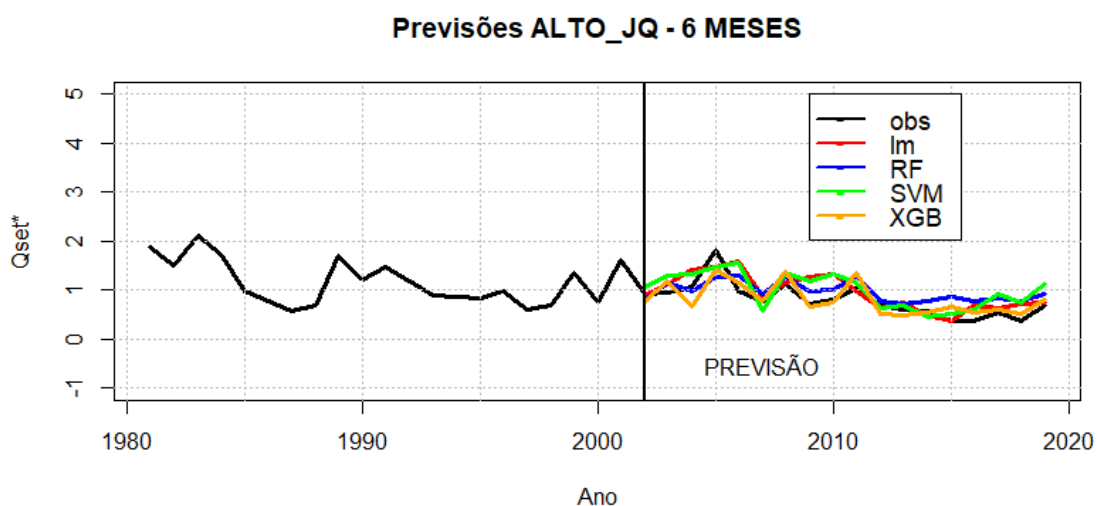
Nota: Entre parêntesis a acurácia apresentada na amostra de teste, fora do parêntesis a acurácia na fase de treino, entre colchetes o viés calculado na amostra de teste

Figura 31-Previsão de vazão Exp2 (Q_{SET}^*) com 6 meses de antecedência bacia ALTO_SF com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). . Preditores utilizados: P_{ONDJFM}^* , Q_{ONDJFM}^* , $PDSI_{ONDJFM}$, VCI_{ONDJFM} , TCI_{ONDJFM}



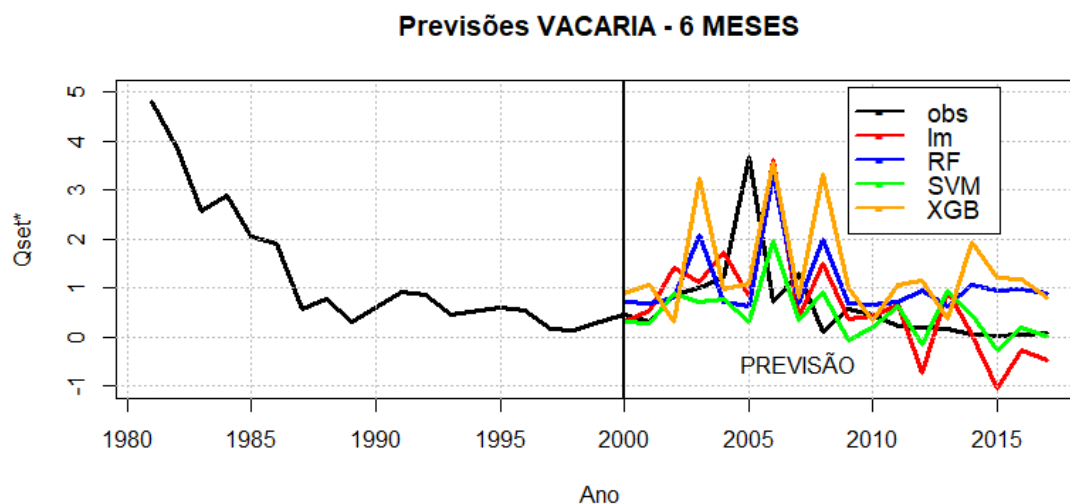
Fonte: Elaborado pelo autor.

Figura 32-Previsão de vazão Exp2 (Q_{SET}^*) com 6 meses de antecedência bacia ALTO_JQ com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{ONDJFM}^* , Q_{ONDJFM}^* , $PDSI_{ONDJFM}$, VCI_{ONDJFM} , TCI_{ONDJFM}



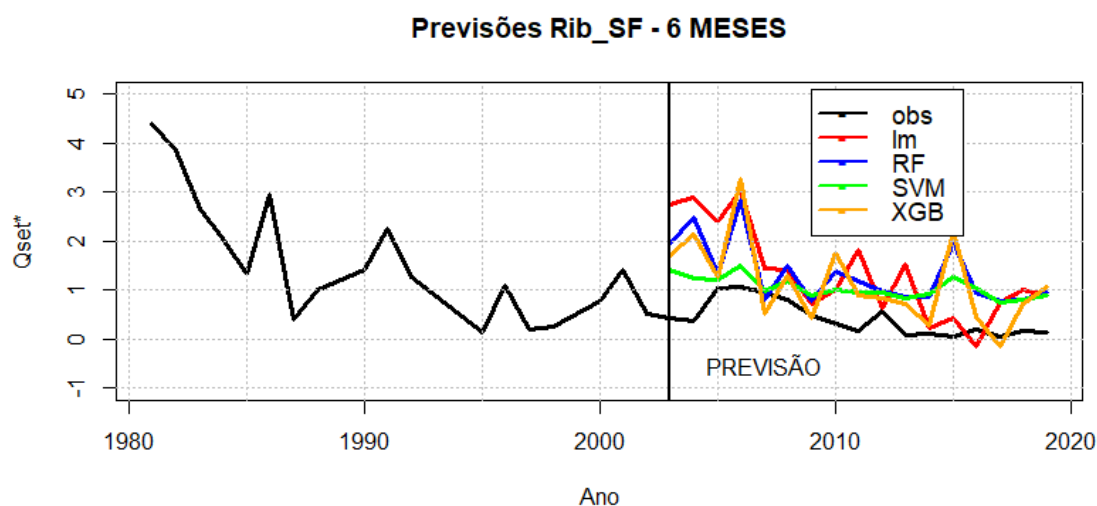
Fonte: Elaborado pelo autor.

Figura 33-Previsão de vazão Exp2 (Q_{SET}^*) com 6 meses de antecedência bacia VACARIA com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{ONDJFM}^* , Q_{ONDJFM}^* , $PDSI_{ONDJFM}$, VCI_{ONDJFM} , TCI_{ONDJFM}



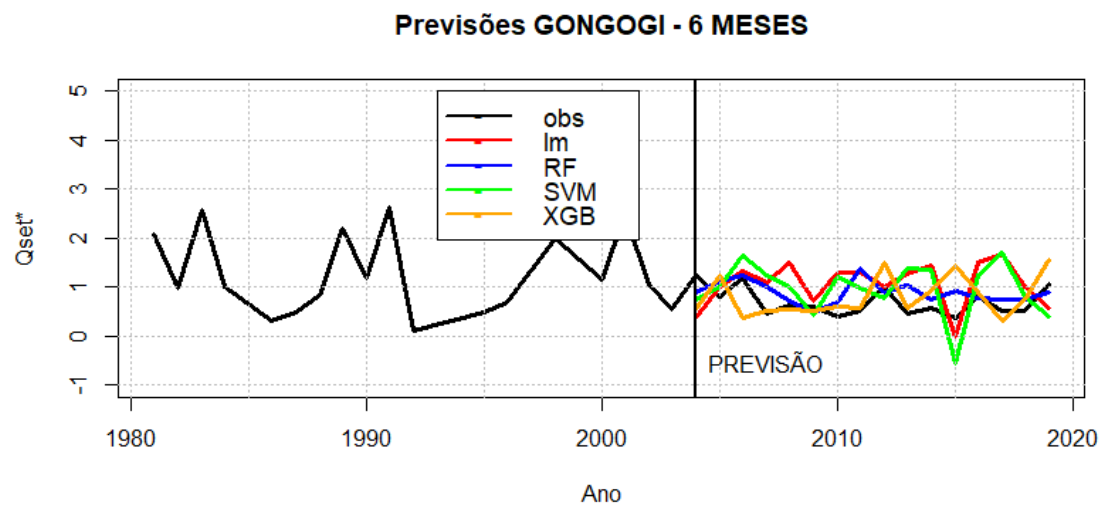
Fonte: Elaborado pelo autor.

Figura 34-Previsão de vazão Exp2 (Q_{SET}^*) com 6 meses de antecedência bacia Rib_SF com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{ONDJFM}^* , Q_{ONDJFM}^* , $PDSI_{ONDJFM}$, VCI_{ONDJFM} , TCI_{ONDJFM}



Fonte: Elaborado pelo autor.

Figura 35-Previsão de vazão Exp2 (Q_{SET}^*) com 6 meses de antecedência bacia GONGOGI com os modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB). Preditores utilizados: P_{ONDJFM}^* , Q_{ONDJFM}^* , $PDSI_{ONDJFM}$, VCI_{ONDJFM} , TCI_{ONDJFM}



Fonte: Elaborado pelo autor.

4.1.3.3 Experimento Exp3 – previsões utilizando os fatores obtidos através das EOFs

O experimento Exp3 teve como objetivo avaliar o ganho de habilidade nas previsões através do emprego das funções empíricas ortogonais (EOFs) para extrair os padrões modais e os fatores de maior variabilidade na bacia de maior extensão territorial do experimento, o ALTO_SF.

Conforme apresentado na **Tabela 13**, os três primeiros fatores explicaram 91,7% da variabilidade da precipitação; os três primeiros fatores explicaram 77,1% da variabilidade do índice TCI; os 4 primeiros fatores explicaram 69,6% da variabilidade do VCI e os dois primeiros fatores explicaram 98,8% do índice Palmer (PDSI).

Os padrões modais dos índices são apresentados na **Figura 37**, onde é possível observar que exceto pelo índice VCI, que não apresentou um padrão espacial bem definido, os demais índices P, TCI e Palmer apresentaram o mesmo padrão espacial, com a diferença de que a precipitação P e o TCI apresentaram um modo a mais na região sudeste (EOF₃) em relação ao PDSI.

Os resultados da aplicação dos modelos ML (lm, RF, SVM-linear e XGBoost) são apresentados na **Tabela 4.5** e analisando os resultados conjuntamente com a **Figura 36**, observa-se que os algoritmos SVM e XGBoost apresentaram as melhores performances, entretanto a performance do modelo extraído pelo SVM foi levemente superior ao XGBoost.

Tabela 13-Variância explicada pelas funções empíricas ortogonais das variáveis predictoras: precipitação (CHIRPS), TCI, VCI, PDSI

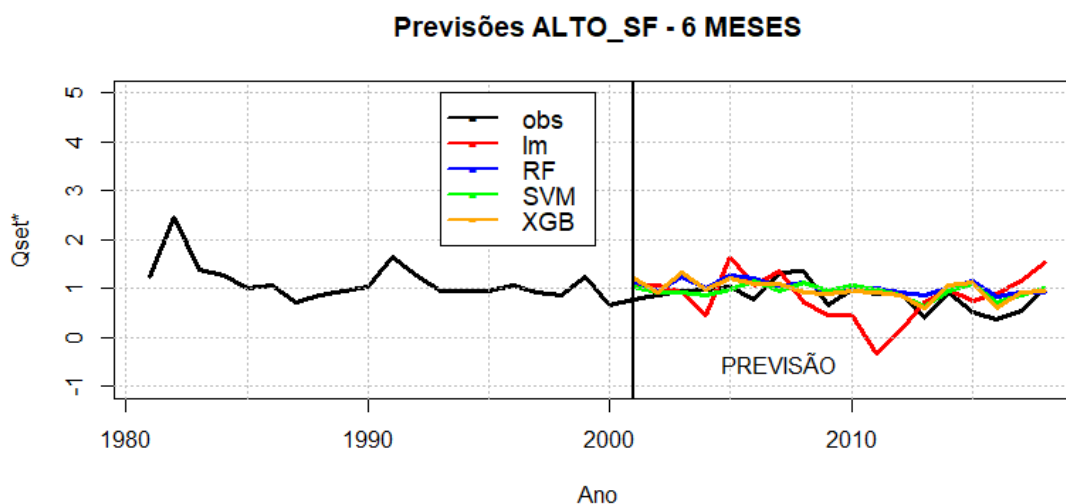
Variável	Variância explicada pelas EOF's	TOTAL
Precipitação CHIRPS	EOF ₁ (32,7%); EOF ₂ (32,1%); EOF ₃ (26,9%)	91,7%
TCI	EOF ₁ (33,1%); EOF ₂ (25,3%); EOF ₃ (18,7%)	77,1%
VCI	EOF ₁ (23,6%); EOF ₂ (17,8%); EOF ₃ (15,4%); EOF ₄ (12,8%)	69,6%
PDSI	EOF ₁ (51,5%); EOF ₂ (47,3%)	98,8%

Tabela 14-Previsão de vazão Exp3 (Q_{SET}^*) com 6 meses de antecedência bacia do ALTO_SF. Os preditores são as EOF's dos campos de precipitação CHIRPS, TCI, VCI, PDSI. Previsões com uso de modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)

BACIA	Treino	Teste	lm	RF	SVM	XGBoost
ALTO_SF	1981-2000	2001-2018	0,079 (0,514) [-0,013]	0,200 (0,307) [+0,189]	0,296 (0,250) [+0,099]	0 (0,28) [+0,132]

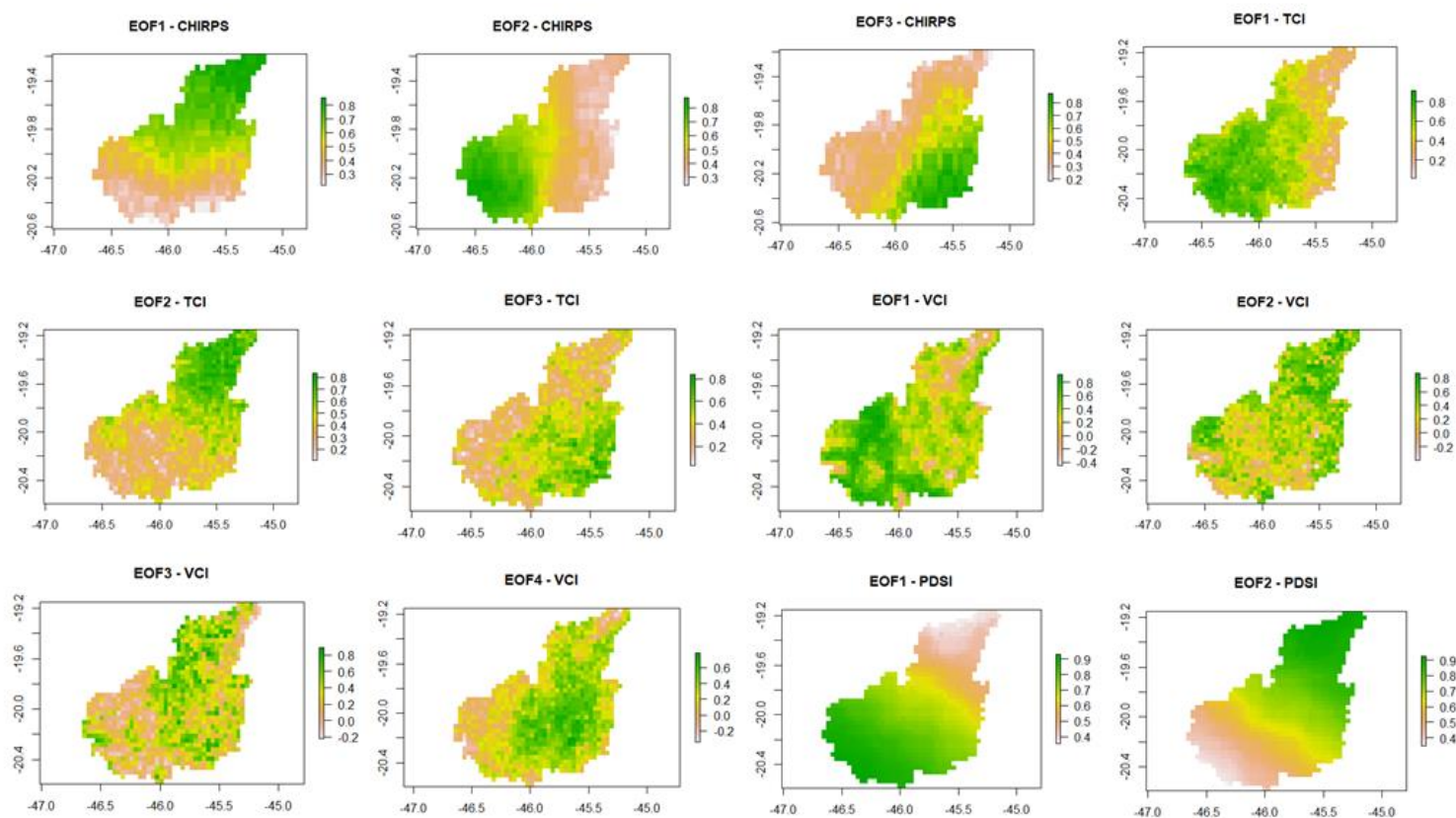
Nota: Entre parêntesis a acurácia RMSE apresentada na amostra de teste, fora do parêntesis a acurácia na fase de treino, entre colchetes o viés calculado na amostra de teste

Figura 36-Previsão de vazão Exp3 (Q_{SET}^*) com 6 meses de antecedência bacia do ALTO_SF. Os preditores são as EOF's dos campos de precipitação CHIRPS, TCI, VCI, PDSI. Previsões com uso de modelos ML: linear (lm), random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)



Fonte: Elaborado pelo autor.

Figura 37- Padrões modais das funções empíricas ortogonais EOFs da precipitação (CHIRPS), TCI, VCI e índice Palmer (PDSI) para a bacia do Alto São Francisco ALTO_SF



Fonte: Elaborado pelo autor.

4.1.3.4 Experimento Exp4 – previsões considerando tendências, influências passadas e remoção de não-estacionariedade

Baseando nas características não-estacionárias do sinal de Q_{SET}^* observadas na **Figura 25** este experimento incluiu as componentes do modelo ARIMA (*Autoregressive Integrated Moving Average*) que se trata de um modelo estatístico explícito amplamente utilizado para realização de previsões em processos não-estacionários (COGHLAN, 2016).

As componentes da formulação modelo ARIMA foram utilizadas para: capturar a valor sem ruído e detectar a sazonalidade pela média móvel (MA), remover não-estacionariedade (I), e considerar correlação entre valores sucessivos (AR) foram consideradas como preditores nesse experimento.

Considerando que o objetivo não é a construção de um modelo autorregressivo, e sim melhorar a habilidade dos modelos ML, a obtenção dos preditores não seguiu rigorosamente os critérios de parcimônia na seleção do melhor candidato e que ao mesmo tempo apresente menos parâmetros.

Baseando-se nos resultados das funções ACF e PACF calculadas para Q_{SET}^* (**Figura 38**) foram incluídas na base de preditores: (1) a média móvel de 3 anos de Q_{ONDJFM}^* e Q_{SET}^* (Q.SMA.3); (2) os valores de Q_{ONDJFM}^{*t-1} , Q_{SET}^{*t-1} ; (3) o valor da derivada de primeira ordem da série suavizada (sem ruído) para melhor capturar a tendência de longo prazo (Q.SMA.3.diff).

Os resultados das previsões com 6 meses de antecedência são apresentados na **Tabela 15** para os algoritmos RF, SVM-Linear e XGBoost. Neste experimento o modelo linear (LM) foi desconsiderado por ter apresentado problemas de degeneração da solução devido à utilização de colunas auto correlacionadas.

Os resultados de RMSE e viés apresentados na **Tabela 15** e os gráficos **Figura 39** a **Figura 43**, mostraram que o modelo obtido pelo algoritmo SVM-Linear superou drasticamente em performance.

O grau de relevância (importância dos preditores) é apresentado na **Tabela 16** onde é possível observar que os preditores obtidos a partir da média móvel (SMA.3) ou de sua derivada (SMA.3.diff) se situaram dentre os mais importantes em todas as bacias. A variável Precipitação esteve entre as mais importantes nas bacias ALTO_SF e GONGOGI. A temperatura TCI esteve entre os 5 preditores mais importantes em todas

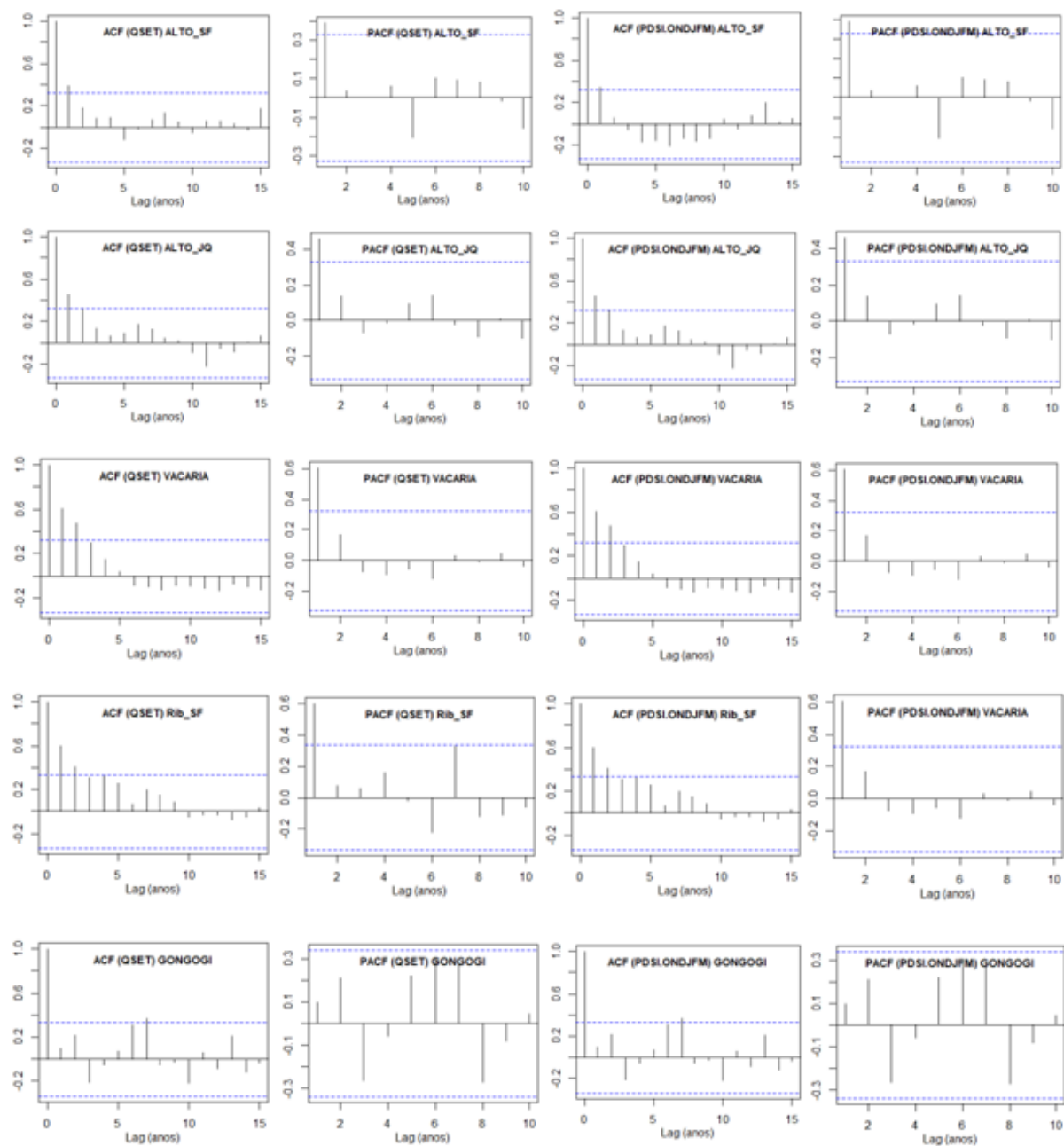
as bacias exceto na bacia do VACARIA. E o PDSI esteve entre as mais importantes somente no ALTO_JQ.

Tabela 15-Previsão de vazão Exp4 (Q_{SET}^*) com 6 meses de antecedência. Os preditores são $TCI_{PONDJFM}^*$, $VCI_{PONDJFM}^*$, $PDSI_{PONDJFM}^*$, P_{ONDJFM}^* , Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{SMA}$, $Q_{SET}^*_{SMA.3.diff}$. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)

BACIA	Treino	Previsão (validação)	RF	SVM	XGBoost
ALTO_SF	1985-2004	2005-2018	0,083 (0,240) [+0,143]	0,048 (0,183) [+0,09]	0,000 (0,213) [+0,100]
ALTO_JQ	1985-2005	2006-2019	0,141 (0,330) [+0,212]	0,033 (0,051) [-0,001]	0,000 (0,417) [+0,004]
VACARIA	1984-2002	2003-2017	0,199 (0,767) [0,000]	0,098 (0,490) [+0,001]	0,001 (1,034) [-0,072]
Rib_SF	1984-2005	2006-2019	0,298 (0,470) [+0,384]	0,442 [0,225] [+0,059]	0,000 [0,510] [+0,346]
GONGOGI	1984-2006	2007-2019	0,175 (0,295) [+0,056]	0,097 (0,207) [+0,151]	0,000 (0,518) [+0,425]

Nota: Entre parêntesis a acurácia RMSE apresentada na amostra de teste, fora do parêntesis a acurácia na fase de treino, entre colchetes o viés calculado na amostra de teste

Figura 38-Autocorrelação (ACF) e autocorrelação parcial (PACF) das séries temporais anuais de QSET* e PDSI_{ONDJFM}



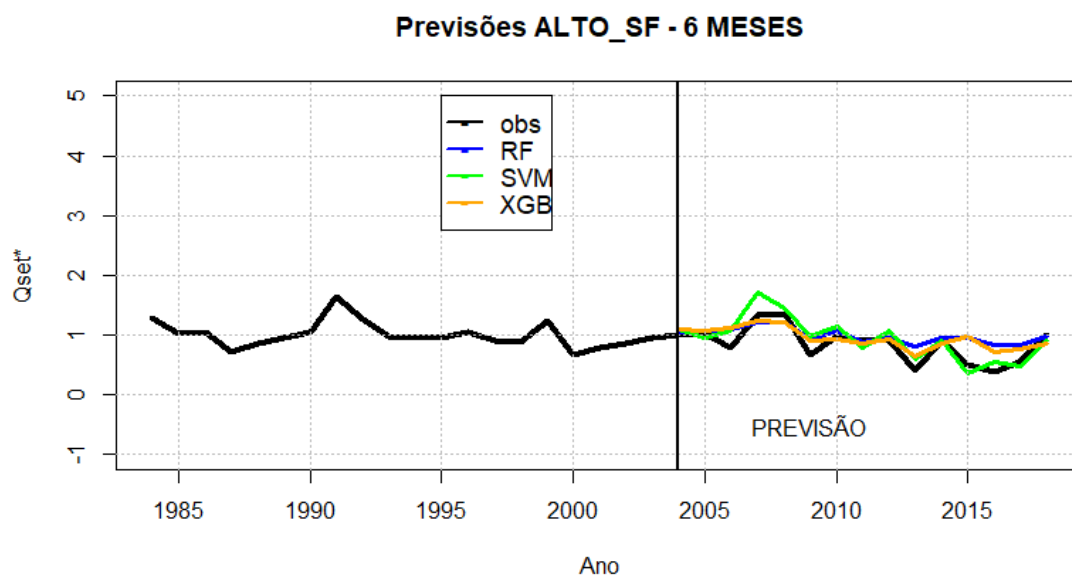
Fonte: Elaborado pelo autor

Tabela 16-Importância das variáveis na previsão de 6 meses de antecedência segundo o modelo que apresentou melhor performance SVM-linear no Exp4

ALTO_SF		ALTO_JQ		VACARIA		RIB_SF		GONGOGI	
QSET.SMA.3.diff	100	QSET.SMA.3	100	QSET.SMA.3	100	QSET.SMA.3	100	QSET.SMA.3.diff	100
Q. ONDJFM	98,84	PDSI.ONDJFM	95,95	Qset_t-1	76,04	TCI.ONDJFM	89,33	P.ONDJFM	81,509
P.ONDJFM	85,53	QSET.SMA.3.diff	79,21	Q.ONDJFM	45,59	VCI.ONDJFM	67,95	TCI.ONDJFM	80,695
QSET.SMA.3	58,86	TCI.ONDJFM	38,55	QONDJFM_t-1	36,47	QSET.SMA.3.diff	67,44	Qset_t-2	65,527
TCI.ONDJFM	53,45	VCI.ONDJFM	33,1	PDSI.ONDJFM	32,68	QONDJFM_t-1	65,05	QSET.SMA.3	62,107
PDSI.ONDJFM	33,14	QONDJFM_t-1	24,72	TCI.ONDJFM	27,25	Q.ONDJFM	51,49	Q.ONDJFM	42,668
VCI.ONDJFM	30,54	Q.ONDJFM	24,16	P.ONDJFM	26,4	P.ONDJFM	40,55	Qset_t-3	34,792
Qset_t-1	17,64	P.ONDJFM	23,91	QSET.SMA.3.diff	23,78	Qset_t-2	22,19	PDSI.ONDJFM	11,556
QONDJFM_t-1	0	Qset_t-1	0	VCI.ONDJFM	0	Qset_t-1	20,34	Qset_t-1	10,101
-	-	-	-	-	-	PDSI.ONDJFM	0	QONDJFM_t-1	3,807
-	-	-	-	-	-	-	-	VCI.ONDJFM	0

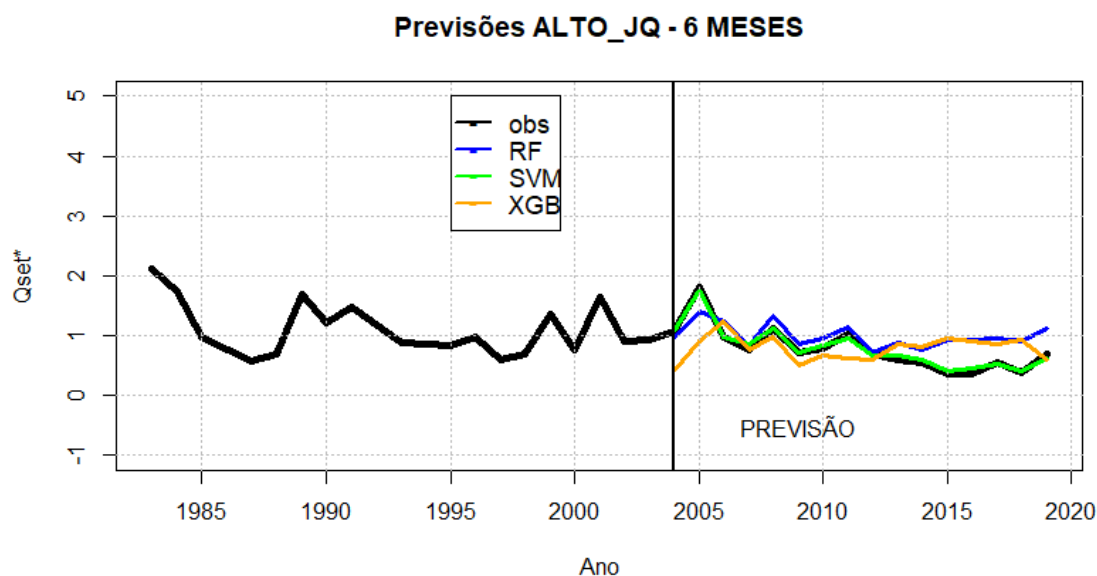
Fonte: elaborado pelo autor

Figura 39-Previsão de vazão Exp4 (QSET*) com 6 meses de antecedência bacia do ALTO_SF
Os preditores são TCI PONDJFM*, VCI PONDJFM*, PDSI PONDJFM*, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_SMA, QSET*_SMA.3.diff. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)



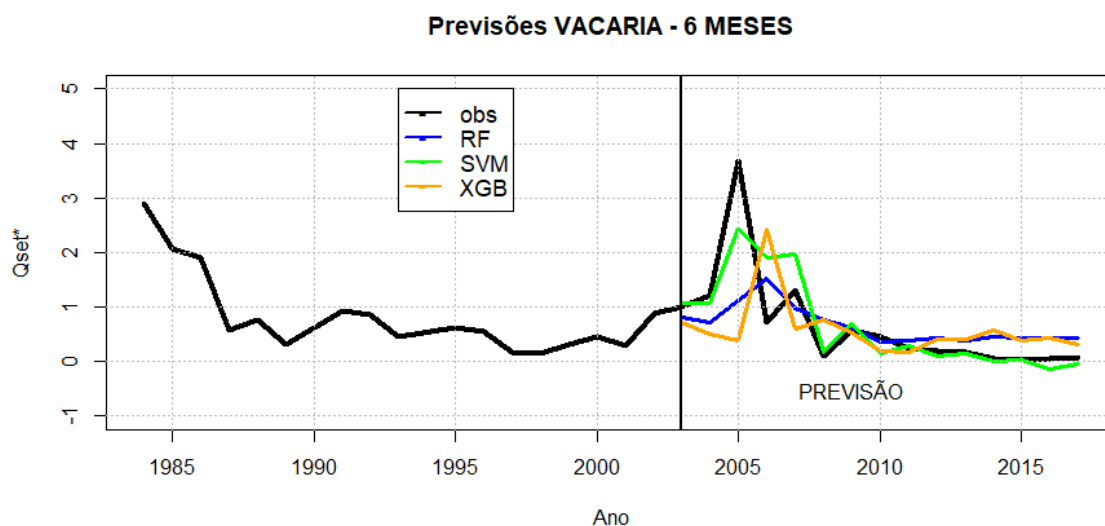
Fonte: Elaborado pelo autor.

Figura 40-Previsão de vazão Exp4 (QSET*) com 6 meses de antecedência bacia do ALTO_JQ.
Os preditores são TCI PONDJFM*, VCI PONDJFM*, PDSI PONDJFM*, PONDJFM*, QONDJFM*, QONDJFM*_t-1, QSET*_t-1, QSET*_SMA, QSET*_SMA.3.diff. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)



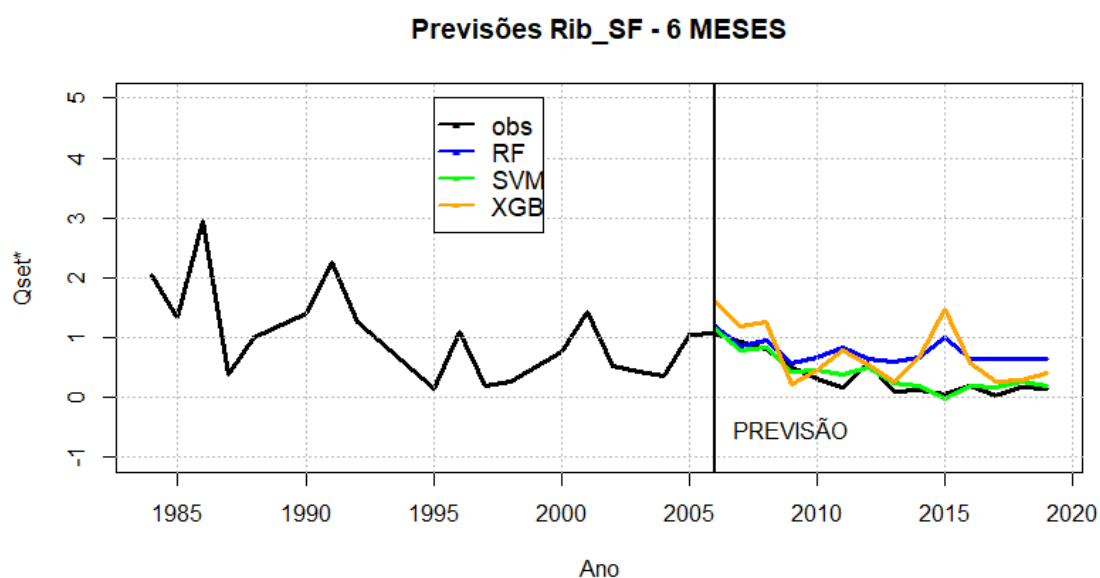
Fonte: Elaborado pelo autor.

Figura 41-Previsão de vazão Exp4 (Q_{SET}^*) com 6 meses de antecedência bacia do VACARIA. Os preditores são $TCI_{PONDJFM}^*$, $VCI_{PONDJFM}^*$, $PDSI_{PONDJFM}^*$, P_{ONDJFM}^* , Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{SMA}$, $Q_{SET}^*_{SMA.3.diff}$. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)



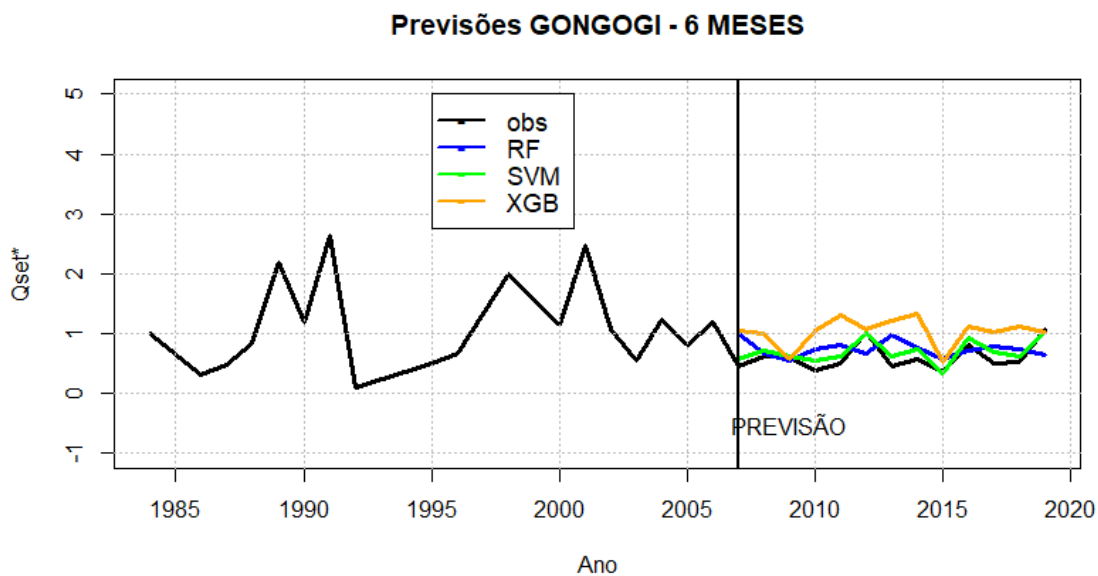
Fonte: Elaborado pelo autor.

Figura 42-Previsão de vazão Exp4 (Q_{SET}^*) com 6 meses de antecedência bacia do Rib_SF. Os preditores são $TCI_{PONDJFM}^*$, $VCI_{PONDJFM}^*$, $PDSI_{PONDJFM}^*$, P_{ONDJFM}^* , Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{SMA}$, $Q_{SET}^*_{SMA.3.diff}$. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)



Fonte: Elaborado pelo autor.

Figura 43-Previsão de vazão Exp4 (Q_{SET}^*) com 6 meses de antecedência bacia do GONGOGI. Os preditores são $TCI_{PONDJFM}^*$, $VCI_{PONDJFM}^*$, $PDSI_{PONDJFM}^*$, P_{ONDJFM}^* , Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{SMA}$, $Q_{SET}^*_{SMA.3.diff}$. Com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB)



Fonte: Elaborado pelo autor.

4.1.3.1 Experimento Exp5 – PDSI como um substituto da Q_{SET}^*

Esse experimento Exp5 testa a habilidade do índice Palmer (PDSI) como um substituto da vazão. Analisando-se a **Tabela 17** juntamente com os gráficos das **Figura 44** a **Figura 48**, observa-se que somente em duas bacias os resultados foram considerados satisfatórios: ALTO_SF com o modelo XGBoost e ALTO_JQ com o modelo SVM.

Mesmo na ausência de dados de vazão observada, a **Tabela 18** mostra que outros índices ganharam maior relevância que o PDSI.

Tabela 17-Previsão de vazão Exp5 (Q_{SET}^*) com 6 meses de antecedência. Os preditores são TCI_{ONDJFM}^* , VCI_{ONDJFM}^* , P_{ONDJFM}^* , $PDSI_{ONDJFM}$, $PDSI_{t-1}$, $PDSI_{ONDJFM_{t-2}}$, $PDSI_{ONDJFM_{t-3}}$, $PDSI_{ONDJFM_SMA.3}$, $PDSI_{ONDJFM_SMA.3.diff}$, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).

BACIA	Treino	Previsão	RF	SVM	XGBoost
ALTO_SF	1985-2004	2005-2018	0,113 (0,283) +0,146	0,153 (0,267) +0,120	0,004 (0,272) +0,099
ALTO_JQ	1985-2005	2006-2019	0,161 (0,431) +0,294	0,273 (0,230) +0,098	0,000 (0,320) +0,127
VACARIA	1984-2002	2003-2017	0,408 (1,002) +0,203	0,621 (1,072) +0,028	0,020 (1,153) +1,102
Rib_SF	1984-2005	2006-2019	0,305 (0,614) +0,548	0,689 (0,568) +0,464	0,138 (0,912) +0,804
GONGOI	1984-2006	2007-2019	0,206 (0,317) (+0,061)	0,326 (0,500) (+0,282)	0,000 (0,326) (0,014)

Nota: Entre parêntesis a acurácia RMSE apresentada na amostra de teste, fora do parêntesis a acurácia na fase de treino, entre colchetes o viés calculado na amostra de teste

Tabela 18-Importância das variáveis na previsão de 6 meses de antecedência segundo o modelo que apresentou melhor performance SVM-linear no Exp5

ALTO_SF		ALTO_JQ		VACARIA		RIB_SF		GONGOI	
P.ONDJFM	100	PDSI_t-3	100	PDSI.ONDJFM	100	TCI.ONDJFM	100	P.ONDJFM	100
TCI.ONDJFM	65,4	P.ONDJFM	86,0	TCI.ONDJFM	89	VCI.ONDJFM	76,1	TCI.ONDJFM	99,1
PDSI.ONDJFM	43,5	TCI.ONDJFM	84,4	P.ONDJFM	87,3	PDSI.diff	75,2	PDSI_t-1	64,3
VCI.ONDJFM	40,7	PDSI.ONDJFM	78,1	PDSI_t-3	75,6	P.ONDJFM	45,4	PDSI.diff	31,3
PDSI.diff	31,7	PDSI.SMA.3	59,0	PDSI_t-1	41,0	PDSI.SMA.3	35,0	PDSI.SMA.3	25,9
PDSI_t-1	21,7	VCI.ONDJFM	51,0	VCI.ONDJFM	33,8	PDSI_t-3	33,9	PDSI.ONDJFM	21,9
PDSI.SMA.3	20,4	PDSI.diff	33,7	PDSI_t-2	32,6	PDSI_t-1	16,3	VCI.ONDJFM	9,0
PDSI_lag3	3,0	PDSI_t-2	15,1	PDSI.SMA.3	32,1	PDSI_t-2	9,7	PDSI_t-3	4,5
PDSI_t-2	0	PDSI_t-1	0	PDSI.diff	0	PDSI.ONDJFM	0	PDSI_t-2	0

Fonte: Elaborado pelo autor

Figura 44-Previsão de vazão Exp5(Q_{SET}^*) com 6 meses de antecedência bacia do ALTO_SF. Os preditores são TCI_{ONDJFM}^* , VCI_{ONDJFM}^* , P_{ONDJFM}^* , $PDSI_{ONDJFM}$, $PDSI_{t-1}$, $PDSI_{ONDJFM_{t-2}}$, $PDSI_{ONDJFM_{t-3}}$, $PDSI_{ONDJFM_SMA.3}$, $PDSI_{ONDJFM_SMA.3.diff}$, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).

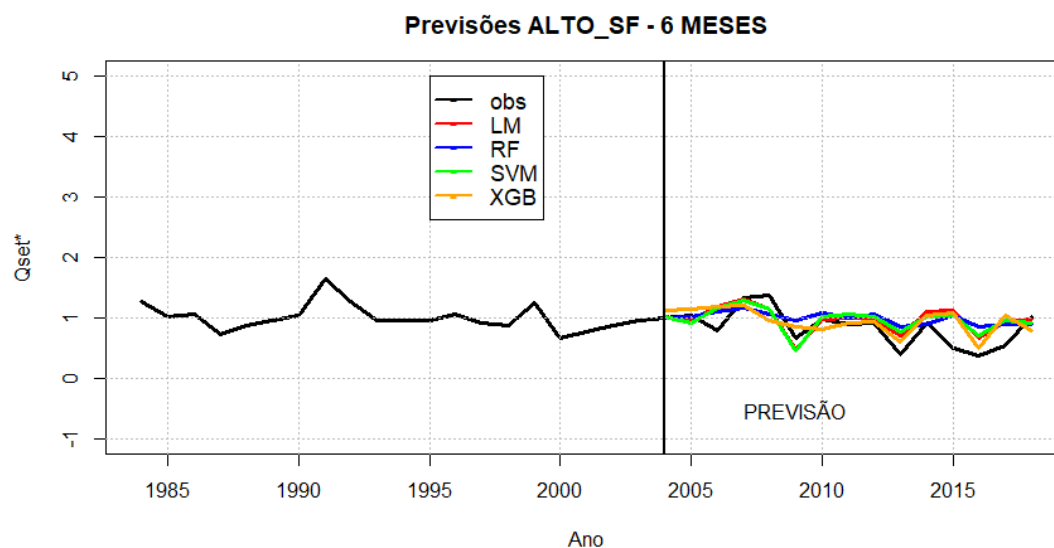


Figura 45-Previsão de vazão Exp5(Q_{SET}^*) com 6 meses de antecedência bacia do ALTO_JQ. Os preditores são TCI_{ONDJFM}^* , VCI_{ONDJFM}^* , P_{ONDJFM}^* , $PDSI_{ONDJFM}$, $PDSI_{t-1}$, $PDSI_{ONDJFM_{t-2}}$, $PDSI_{ONDJFM_{t-3}}$, $PDSI_{ONDJFM_SMA.3}$, $PDSI_{ONDJFM_SMA.3.diff}$, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).

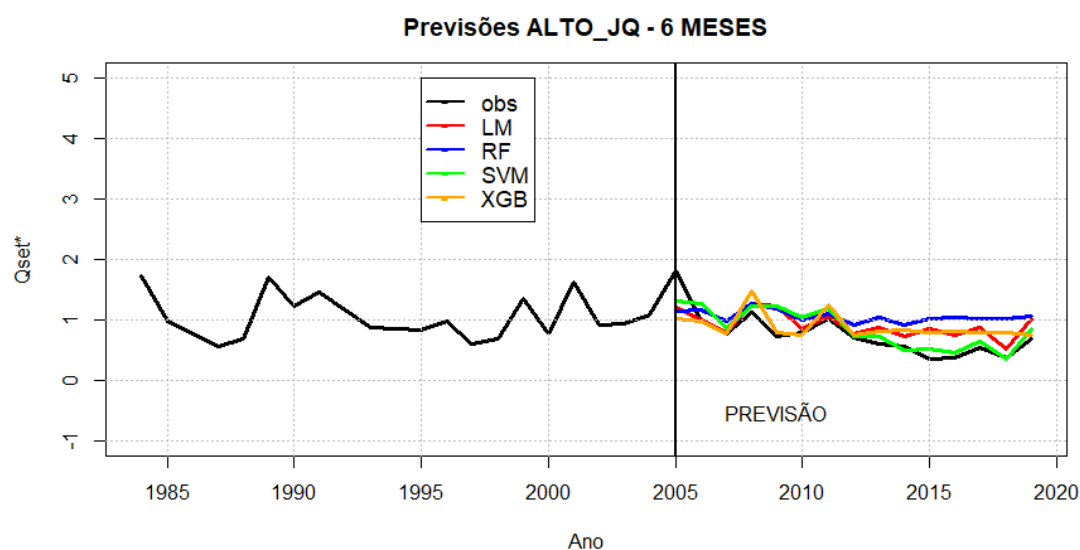
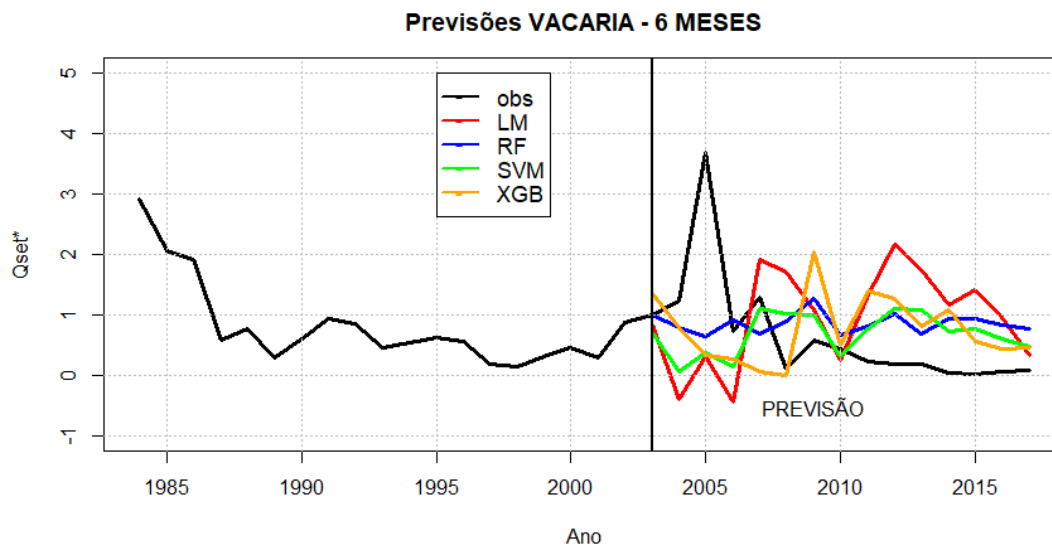
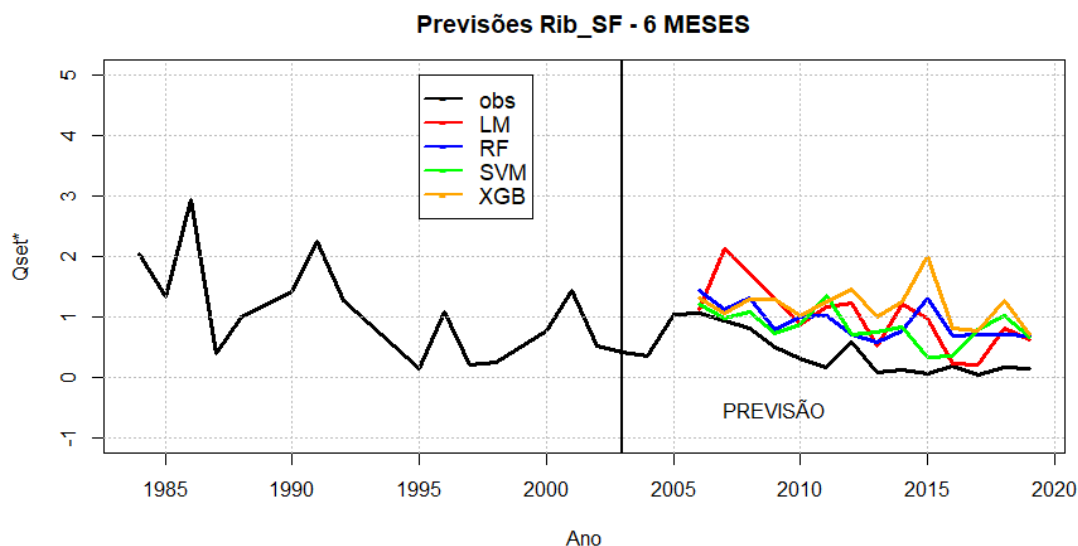


Figura 46-Previsão de vazão Exp5(Q_{SET}^*) com 6 meses de antecedência bacia do VACARIA. Os preditores são TCI_{ONDJFM}^* , VCI_{ONDJFM}^* , P_{ONDJFM}^* , $PDSI_{ONDJFM}$, $PDSI_{t-1}$, $PDSI_{ONDJFM_{t-2}}$, $PDSI_{ONDJFM_{t-3}}$, $PDSI_{ONDJFM_SMA.3}$, $PDSI_{ONDJFM_SMA.3.diff}$, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).



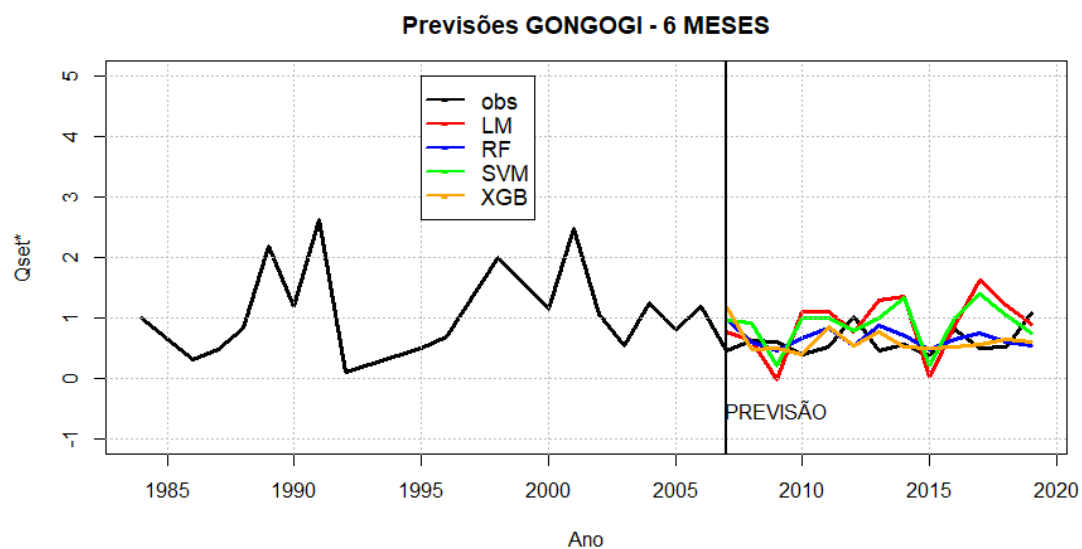
Fonte: Elaborado pelo autor.

Figura 47-Previsão de vazão Exp5(Q_{SET}^*) com 6 meses de antecedência bacia do Rib_SF. Os preditores são TCI_{ONDJFM}^* , VCI_{ONDJFM}^* , P_{ONDJFM}^* , $PDSI_{ONDJFM}$, $PDSI_{t-1}$, $PDSI_{ONDJFM_{t-2}}$, $PDSI_{ONDJFM_{t-3}}$, $PDSI_{ONDJFM_SMA.3}$, $PDSI_{ONDJFM_SMA.3.diff}$, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).



Fonte: Elaborado pelo autor.

Figura 48-Previsão de vazão Exp5(Q_{SET}^*) com 6 meses de antecedência bacia do GONGOI. Os preditores são TCI_{ONDJFM}^* , VCI_{ONDJFM}^* , P_{ONDJFM}^* , $PDSI_{ONDJFM}$, $PDSI_{t-1}$, $PDSI_{ONDJFM_{t-2}}$, $PDSI_{ONDJFM_{t-3}}$, $PDSI_{ONDJFM_SMA.3}$, $PDSI_{ONDJFM_SMA.3.diff}$, com uso de modelos ML: random forest (RF), máquina de vetor de suporte (SVM linear), XGBoost (XGB).



Fonte: Elaborado pelo autor.

4.1.3.2 Experimento Exp6 – Teste de generalização em bacia vizinha com dados locais – transferência de modelo ML SVM

O experimento Exp6 testou a capacidade de generalização do modelo SVM-Linear, que apresentou melhor performance nos experimentos anteriores, treinado em uma bacia distinta, em realizar previsões de vazões mínimas em outra bacia com dados locais (sem transferência de dados hidrológicos). Foram testadas das de forma cruzada: ALTO_JQ $\otimes$ GONGOGI; ALTO_JQ $\otimes$ ALTO_SF e VACARIA $\otimes$ Rib_SF.

Os resultados são apresentados na tabela **Tabela 19**, onde se observa um grau de correlação significativo entre os valores previstos com o modelo exógeno utilizando os preditores calculados com dados locais.

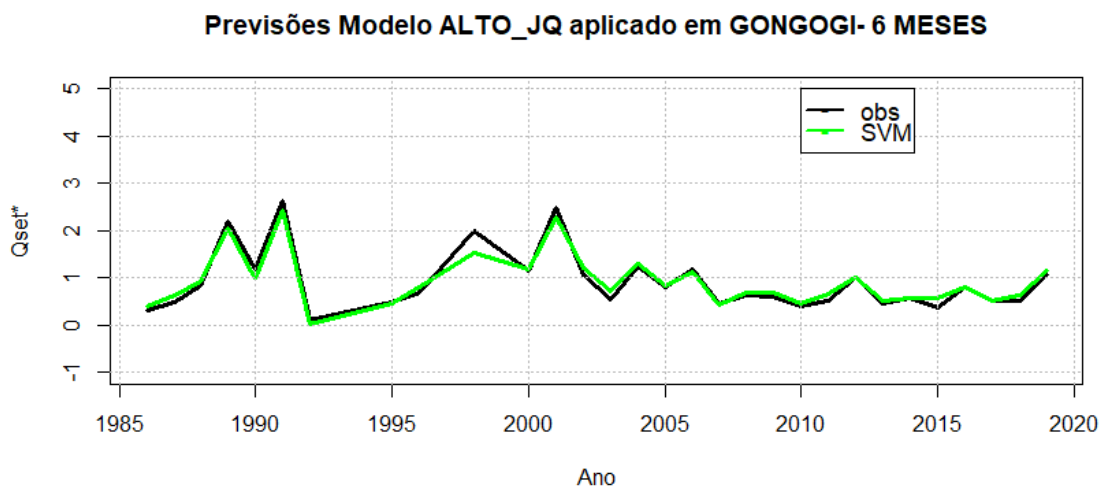
O treinamento envolveu toda a amostra de dados na bacia onde o modelo foi originalmente treinado, e a previsão (validação) envolveu todos os dados da bacia testada. Os resultados do experimento também são apresentados nas **Figura 49** a **Figura 51**, onde se observa que os modelos capturaram pequenas variações mínimas no período mais afetado pela seca. Na bacia do Rib_SF a transferência do modelo chegou a gerar valores levemente negativos.

Tabela 19-Previsão de vazão Exp6 (Q_{SET}^*) com 6 meses de antecedência. Os preditores são TCI_{ONDJFM} , VCI_{ONDJFM} , $PDSI_{ONDJFM}$, P_{ONDJFM}^* , Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET}^*_{t-3}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$, com uso de modelos ML: máquina de vetor de suporte (SVM linear)

Validação Cruzada	Previsão	SVM
ALTO_JQ (treino) $\otimes$ GONGOGI (teste)	1986-2019	0,147 (0,949) [+0,034]
ALTO_JQ (treino) $\otimes$ ALTO_SF (teste)	1985-2018	0,073 (0,950) [-0,038]
VACARIA (treino) $\otimes$ Rib_SF (teste)	1985-2019	0,419 (0,675) [+0,011]

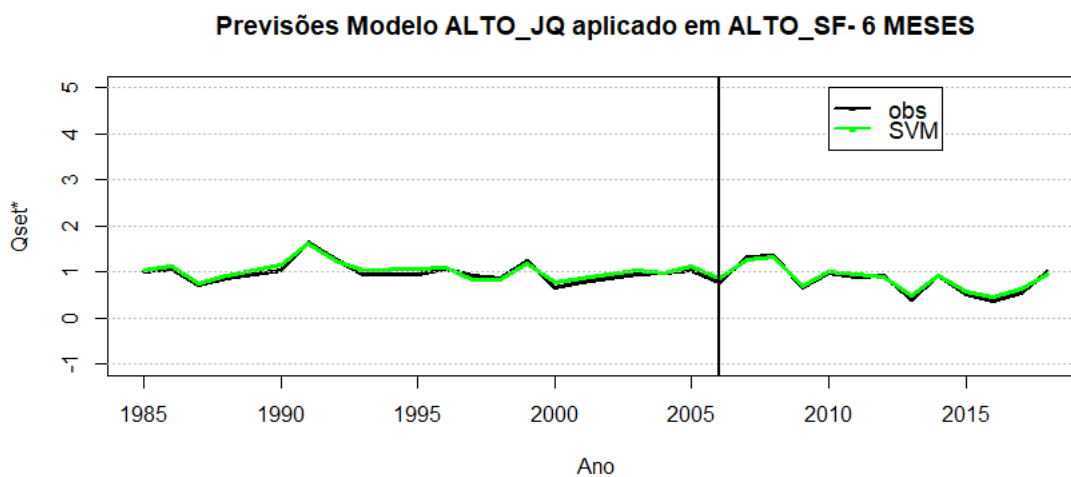
Nota: Entre parêntesis o coeficiente de correlação apresentada na bacia de teste, fora do parêntesis a acurácia RMSE na bacia teste, entre colchetes o viés calculado na bacia de teste

Figura 49-Previsão de vazão cruzada Exp6 aplicando-se o modelo SVM obtido para a bacia do ALTO_JQ para gerar previsões na bacia do rio GONGOGI com dados locais e 6 meses de antecedência.



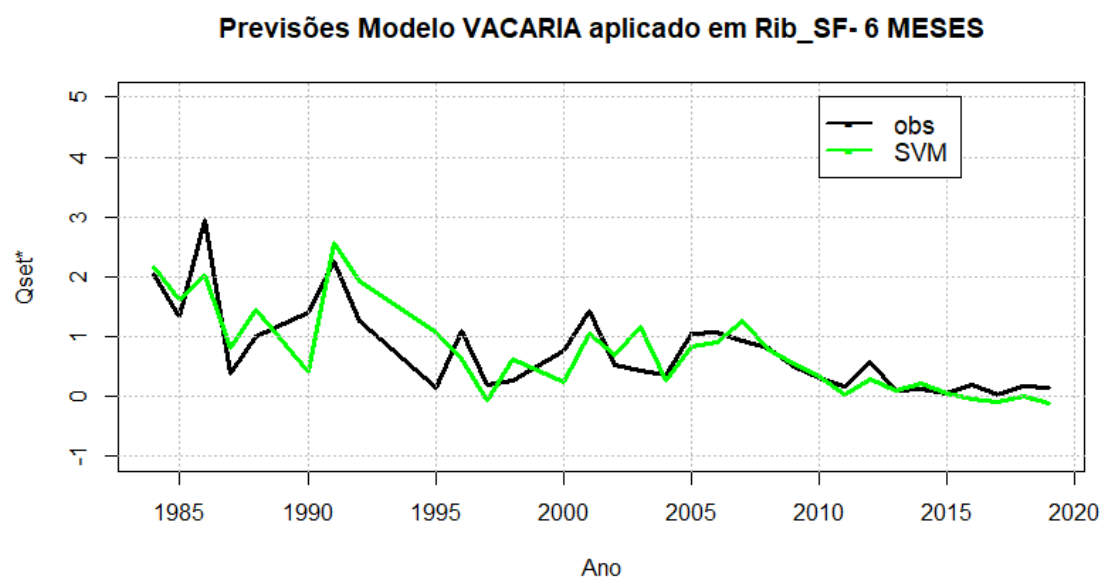
Fonte: Elaborado pelo autor.

Figura 50-Previsão de vazão cruzada Exp6 aplicando-se o modelo SVM obtido para a bacia do ALTO_JQ para gerar previsões de vazão Q_{SET}^* na bacia do rio ALTO_SF com dados locais e 6 meses de antecedência.



Fonte: Elaborado pelo autor.

Figura 51- Previsão de vazão cruzada Exp6 aplicando-se o modelo SVM obtido para a bacia do rio VACARIA para gerar previsões de vazão Q_{SET}^* na bacia do rio Rib_SF com dados locais e 6 meses de antecedência.



Fonte: Elaborado pelo autor.

4.1.3.3 Experimento Exp7 – teste de generalização bacia vizinha sem dados locais

O experimento Exp7 teve como objetivo testar o poder de generalização do modelo SVM-Linear, que apresentou melhor performance nos experimentos anteriores, numa bacia sem dados.

Uma vez que o experimento Exp6 mostrou que os modelos gerados individualmente podem ser generalizados, esse experimento avalia o impacto da transferência de informações em bacias sem informações hidrológicas.

Foram estadas das de forma cruzada: ALTO_JQ $\otimes$ GONGOGI; ALTO_JQ $\otimes$ ALTO_SF e VACARIA $\otimes$ Rib_SF. Os resultados apresentados na **Tabela 20** mostram um coeficiente de correlação baixo ($<0,5$) nos testes ALTO_JQ $\otimes$ GONGOGI e ALTO_JQ $\otimes$ ALTO_SF, entretanto, observa-se nas **Figura 52** a **Figura 54** que a modelagem cruzada em locais sem dados conseguiu acompanhar a tendência de valores mínimos abaixo da média, especialmente nos anos de extrema seca, porém não conseguiu reproduzir satisfatoriamente valores acima da média.

Tabela 20- Previsão de vazão cruzada Exp7 (Q_{SET}^*) com 6 meses de antecedência. Os preditores são TCI_{ONDJFM} , VCI_{ONDJFM} , $PDSI_{ONDJFM}$, P_{ONDJFM}^* , Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$, com uso de modelos ML: máquina de vetor de suporte (SVM linear). As variáveis Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$ são transferidas da bacia com dados onde foi calibrado o modelo.

Validação Cruzada	Previsão	SVM
ALTO_JQ (treino) $\otimes$ GONGOGI (teste)	1986-2019	0,142 (0,451) [-0,089]
ALTO_JQ (treino) $\otimes$ ALTO_SF (teste)	1985-2018	0,142 (0,276) [+0,034]
VACARIA (treino) $\otimes$ Rib_SF (teste)	1985-2019	0,503 (0,602) [+0,068]

Nota: Entre parêntesis o coeficiente de correlação apresentada na bacia de teste, fora do parêntesis a acurácia RMSE na bacia teste, entre colchetes o viés calculado na bacia de teste

Figura 52- Previsão de vazão cruzada Exp7 (Q_{SET}^*) com 6 meses de antecedência na bacia ALTO_SF com modelo e dados da bacia ALTO_JQ. Os preditores são TCI_{ONDJFM} , VCI_{ONDJFM} , $PDSI_{ONDJFM}$, $PONDJFM^*$, $QONDJFM^*$, $QONDJFM^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$, com uso de modelos ML: máquina de vetor de suporte (SVM linear). As variáveis $QONDJFM^*$, $QONDJFM^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$ são transferidas da bacia com dados onde foi calibrado o modelo.

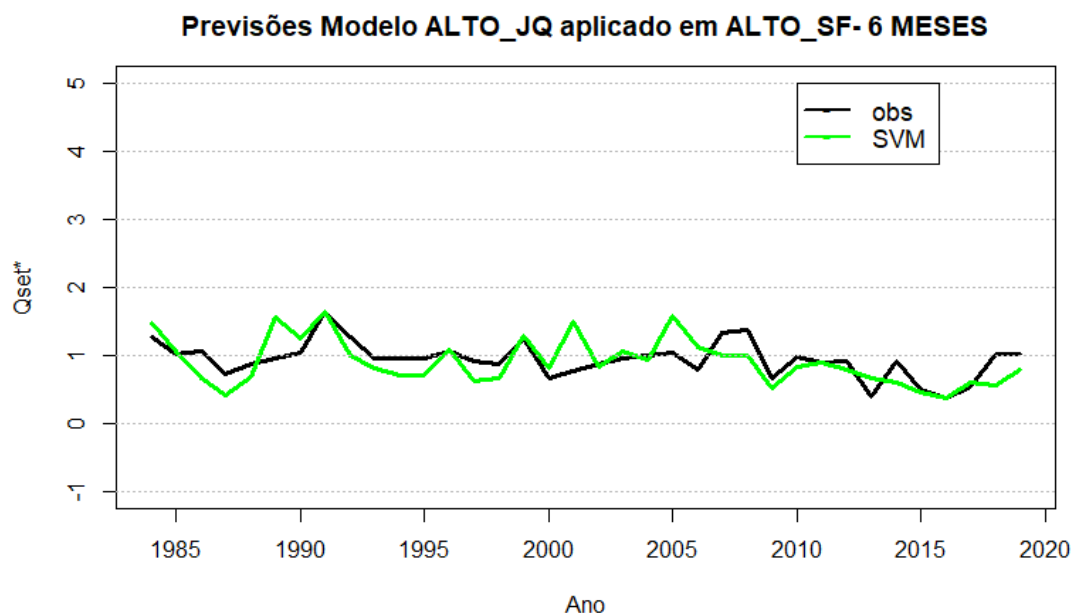


Figura 53- Previsão de vazão cruzada Exp7 (Q_{SET}^*) com 6 meses de antecedência na bacia GONGOGI com modelo e dados da bacia ALTO_JQ. Os preditores são TCI_{ONDJFM} , VCI_{ONDJFM} , $PDSI_{ONDJFM}$, $PONDJFM^*$, $QONDJFM^*$, $QONDJFM^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$, com uso de modelos ML: máquina de vetor de suporte (SVM linear). As variáveis $QONDJFM^*$, $QONDJFM^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$ são transferidas da bacia com dados onde foi calibrado o modelo.

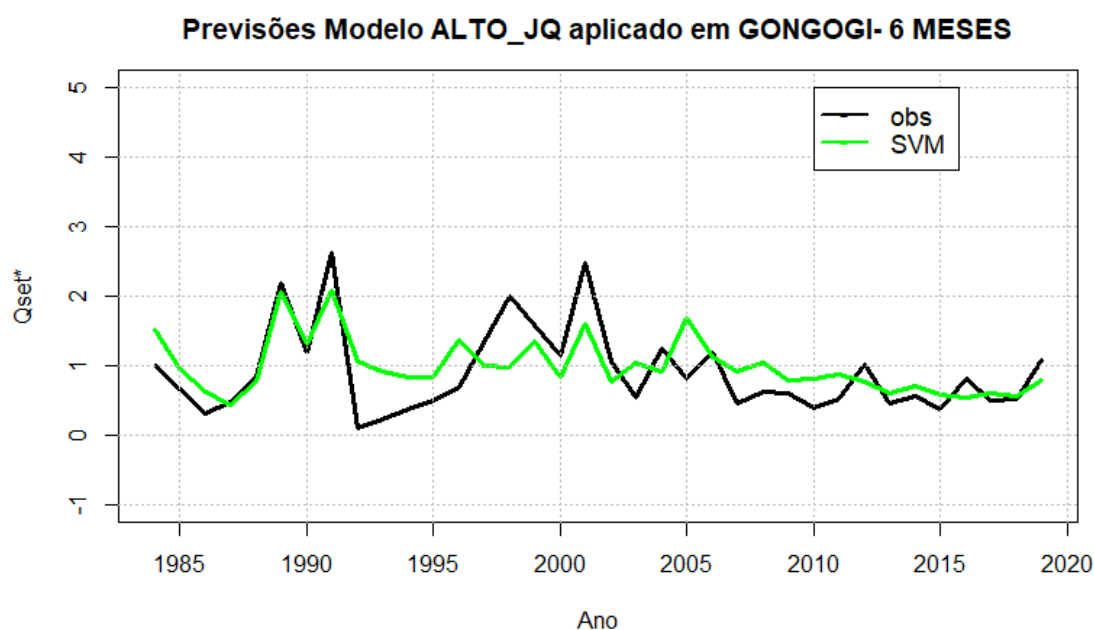
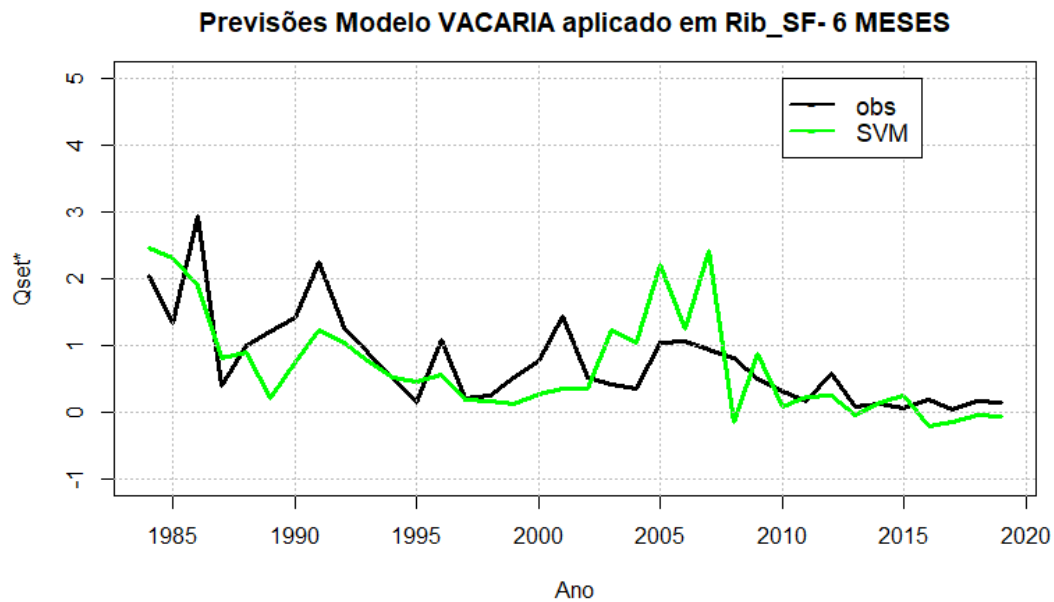


Figura 54- Previsão de vazão cruzada Exp7 (Q_{SET}^*) com 6 meses de antecedência na bacia do rio Rib_SF com modelo e dados da bacia VACARIA. Os preditores são TCI_{ONDJFM} , VCI_{ONDJFM} , $PDSI_{ONDJFM}$, P_{ONDJFM}^* , Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$, com uso de modelos ML: máquina de vetor de suporte (SVM linear). As variáveis Q_{ONDJFM}^* , $Q_{ONDJFM}^*_{t-1}$, $Q_{SET}^*_{t-1}$, $Q_{SET}^*_{t-2}$, $Q_{SET.SMA.3}$, $Q_{SET.SMA.3.diff}$ são transferidas da bacia com dados onde foi calibrado o modelo.



5 DISCUSSÃO

5.1 CARACTERÍSTICAS DO SINAL DA VAZÃO MÍNIMA Q_{SET}

A decomposição do sinal no domínio do tempo e da frequência realizada através da transformada de wavelets (**Figura 24**) demonstrou que a série de vazões mínimas em setembro (Q_{SET}) apresenta fortes características de não-estacionariedade. Após 2010 se verificou uma perda de potência generalizada de sinal em todas as bacias, exceto na bacia ALTO_SF. Os testes de tendência MK e Pettit **Figura 21** também evidenciaram uma mudança de tendência significativa próximo a este ano em quase todas as bacias, porém com diferentes níveis de significância, demonstrando que o evento da seca de 2010-2016, uma das mais severas da história, impactou extremamente o regime de vazões mínimas.

As tendências de redução de precipitação e vazão apresentadas na **Figura 21** na área de experimento deste trabalho também foram identificadas em outras regiões do bioma cerrado no estado de Minas Gerais (JUNQUEIRA et al., 2020b), Tocantins (JUNQUEIRA et al., 2020a) e também na bacia do rio Doce localizada no bioma Mata Atlântica (JESUS et al., 2020), impactando drasticamente as vazões superficiais e as recargas subterrâneas.

JUNQUEIRA et al. (2020b) ainda destacam na sua área de estudo no bioma cerrado, que as vazões de base são grande parte do hidrograma, e que o tempo médio de resposta de uma seca meteorológica para uma seca hidrológica é de 0-6 meses, concluindo que variações na vazão são mais relacionadas com mudanças que ocorrem dentro da mesma estação do que em anos prévios. Porém, nos experimentos realizados por JESUS et al. (2020) os efeitos da seca meteorológica sobre a recarga foram mais demorados e persistidos chegando à escala anual.

O experimento realizado mostrou que a premissa de controle intra-anual da vazão mínima, que pressupõe controle pelos preditores estimados dentro da estação chuvosa dentro do mesmo ano hidrológico (Exp2), geraram previsões pobres para rios com baixos índices pluviométricos e baixa capacidade de recarga (ver Figura 16). Quando se agregam observações de anos anteriores na base de preditores (Exp4) a melhoria na

habilidade de previsão é drasticamente superior. Nossos resultados mostraram influências em escala temporal de 2-3 anos.

A componente de baixa frequência dos sinais de Q_{SET} , exerce grande controle sobre a variabilidade anual das vazões mínimas. Isso fica evidenciado no ganho de habilidade ao se comparar os resultados do experimento Exp 5 com os demais, quando se agrega os sinais suavizados (SMA) e sua respectiva derivada de primeira ordem (SMA.diff) na base de variáveis preditoras. Na **Tabela 18** observa-se que pelo menos uma dessas variáveis SMA e SMA.diff se posiciona em 1º lugar do rank em todos os experimentos.

Fisicamente, a componente SMA de Q_{SET} é representada pela variação das reservas subterrâneas da bacia, e pelo controle exercido pelo nível de armazenamento dos sistemas aquíferos, sobre as vazões mínimas. Tais processos hidrológicos possuem a característica de variações em baixa frequência e com regime de variação interanual, uma vez que são grandes reservatórios com grande variabilidade nas entradas e retiradas.

O processo de variação das reservas subterrâneas que controlam em grande parte a Q_{SET} , é correlacionado às reposições e depleções que ocorrem durante anos, e não pela mudança de estação chuvosa para seca dentro do mesmo ano hidrológico. Isso fica evidenciado nas bacias com menores alturas pluviométricas anuais: Vacaria ($862 \text{ mm}\cdot\text{ano}^{-1}$), Rib_SF ($857 \text{ mm}\cdot\text{ano}^{-1}$) e Gongogi ($866 \text{ mm}\cdot\text{ano}^{-1}$) e com baixa capacidade de recarga (**Figura 16**) nas quais se obteve maiores ganhos de habilidade de previsão devido à inclusão de SMA e SMA.diff na base de preditores. Ao se comparar os experimentos Exp2 com o Exp4, no primeiro observa-se que a performance dos modelos nestas três bacias não foi satisfatória, embora os modelos tenham conseguido capturar a componente anual da variação (**Figura 33, Figura 34 e Figura 35**)

5.2 OS PREDITORES VCI, TCI

O índice VCI se apresentou negativamente correlacionado com a vazão de Q_{SET} e com a precipitação, além de ter demonstrado uma tendência de crescimento significativa em algumas bacias, com aceleração de tendência nos últimos 10 anos. Algumas explicações podem ser dadas para esse efeito associativo. Uma delas é que a evapotranspiração potencial (ETP) é negativamente correlacionada com a precipitação, dado que nos períodos chuvosos a temperatura do ar e a insolação são menores

(COLLISCHONN; TUCCI, 2014) o que se reflete em termos de evapotranspiração real (ET_r) quando se adentra na estação chuvosa no cerrado (FLEISCHMANN et al., 2019).

O próprio idealizador do índice VCI, ressalta maiores preocupações de sua aplicação em situações de excessiva umidade do solo ou presença de nuvens por mais de 3 semanas, uma vez que nessas situações o sinal do vigor da vegetação é consideravelmente reduzido (KOGAN, 1995). Também em ambientes úmidos do bioma cerrado, em que a água não é o fator limitante da dinâmica da vegetação, a temperatura exerce papel fundamental na dinâmica de expansão da vegetação, e algumas áreas de pântanos quando o nível do lençol freático se rebaixa, provocando expansão da vegetação (DOS SANTOS et al., 2020).

Além disso, maior presença de água no sistema radicular das plantas não necessariamente significa em maior infiltração profunda de modo a resultar numa recarga efetiva. Esses fatores explicam o motivo do índice VCI não ter se situado entre os preditores mais importantes (**Tabela 16** do experimento Exp4).

Um outro ponto que merece menção, é que o índice VCI, por ser calculado a partir de valores máximos e mínimos do NDVI, pode se tornar bastante sensível a mudanças nos padrões de uso e ocupação do solo.

Já o índice TCI se destacou entre os cinco mais importantes preditores em quase todas as bacias, uma vez que representa mais diretamente o processo de evapotranspiração. No Cerrado brasileiro a redução da precipitação e aumento da temperatura tem intensificado a ocorrência de secas onde se observou um incremento de 1,3°C no período de 2013-2018 resultando num aumento da taxa de evapotranspiração de 81,4% para 88,4% em sua área de estudo (JUNQUEIRA et al., 2020a, 2020b).

5.3 PALMER (PDSI)

O índice Palmer (PDSI) não apresentou grande relevância no ranking de preditores (**Tabela 16**), apesar da sua correlação significativa com a vazão Q_{SET} (**Figura 22**) e de estudos anteriores terem mostrado forte correlação deste índice com a total escoado anual (DAI, 2011b, 2011a).

Embora seja um índice persistido, as wavelets cruzadas (WTC) apresentadas na **Figura 23** mostraram que a coerência (correlação) entre os sinais de PDSI e Q são bem significativas, entretanto segmentadas no domínio do tempo e frequência. A correlação entre essas duas variáveis na baixa frequência (até 12 meses) existe, mas não é contínua no tempo, o que prejudica a utilização desse índice como preditor.

Fortes correlações entre PDSI e Q foram encontradas na escala de 32 meses (**Figura 23**), entretanto o experimento Exp5 realizado nessa escala temporal não gerou resultados promissores. Do ponto de vista físico, sem a vazão observada Q , o modelo fica impossibilitado de calcular a variação das reservas subterrâneas e, portanto, relacionar Q_{SET} . Ficou evidenciado que o PDSI não possui essa habilidade, uma vez que é um índice relacionado as reservas superficiais e subsuperficiais (WEBER; NKEMDIRIM, 1998).

Nas bacias ALTO_SF e ALTO_JQ, bem supridas por precipitação e com condições de recarga favoráveis, a componente de interanual de variação dos sistemas aquíferos tem menor influência na variação de vazões mínimas Q_{SET} em relação às demais, uma vez que não levam déficits de recarga para o ano seguinte. Nestas, a componente anual aprendida pelos modelos ML levou a um melhor desempenho mesmo quando variáveis representativas de influências passadas não foram consideradas na base de preditores.

5.4 A VARIABILIDADE ESPACIAL

Também, quando se comparam os experimentos Exp3 e Exp2 observa-se que a tentativa de melhoria da acurácia através da decomposição das variáveis em padrões modais (EOFs) e fatores (sinais temporais) agregaram pouca acurácia aos modelos, que pode ser explicado pela escala temporal do horizonte previsão, sendo 6 meses um período demasiadamente longo para que processos hidrológicos localizados pudessem apresentar alguma diferença na vazão prevista. Além disso, os padrões modais de precipitação, temperatura e PDSI mostraram bastante semelhança.

5.5 O QUE OS ALGORITMOS APREENDERAM SOBRE HIDROLOGIA?

Em termos hidrológicos o experimento Exp5 demonstrou habilidade de aprendizado de alguns processos hidrológicos em escala sazonal, anual e interanual envolvidos no cálculo de vazões mínimas, mesmo em situações de extrema seca, período não incluído na amostra de treinamento.

Conforme já mencionado previamente, as vazões mínimas no período de estiagem são dependentes do nível de armazenamento das reservas subterrâneas, que por

sua vez depende do nível de armazenamento no início do ano hidrológico mais a variação ocorrida ao seu término (**Equação 34**).

Conforme ilustrado na **Figura 55**, a variação das reservas no momento da previsão com seis meses de antecedência é dada pela **Equação 35**, e corresponde à variação da reserva após o término da estação chuvosa, com 6 meses de antecedência ao alvo. Conhecendo-se essa parcela de variação, o algoritmo também aprende através dos dados, a relação entre a variação anual e a variação do semestre chuvoso (**Equação 36**).

A partir das tendências, valores anteriores, detecção de sazonalidade e estimativa da variação anual ao término da estação chuvosa, o algoritmo extrai a relação entre o armazenamento ao término do ano hidrológico (S_{SET}) dado pela **Equação 37**. Por fim, deriva a relação entre o armazenamento e a vazão (**Equação 38**).

O aprendizado do processo hidrológico descrito acima é induzido de forma implícita, a partir do treinamento sobre os dados observados. Outras formas de se estimar a vazão e o armazenamento são possíveis, envolvendo sensoriamento remoto fazendo uso do GRACE (CAO et al., 2019; GONÇALVES et al., 2020), modelagem hidrológica ou uso de parâmetros empíricos de literatura, tal qual como na formulação clássica de decaimento exponencial $Q_t = Q_0 \cdot e^{-\alpha t}$, onde α é o coeficiente de depleção, e Q_0 é a vazão no início da recessão.

Trabalhos envolvendo sensoriamento remoto através do GRACE requerem séries mais longas para treinamento de modelos estatísticos e sua baixa resolução espacial pode ser um limitante para seu emprego em pequenas bacias. Já a determinação de um coeficiente de depleção pode se tornar uma tarefa desafiadora. SHARMA; BISWAL (2022), após estudar 408 bacias utilizando modelagem hidrológica para realizar previsões de recessão, chegaram à conclusão de que não existe nenhum valor de α suficientemente representativo para aplicações práticas.

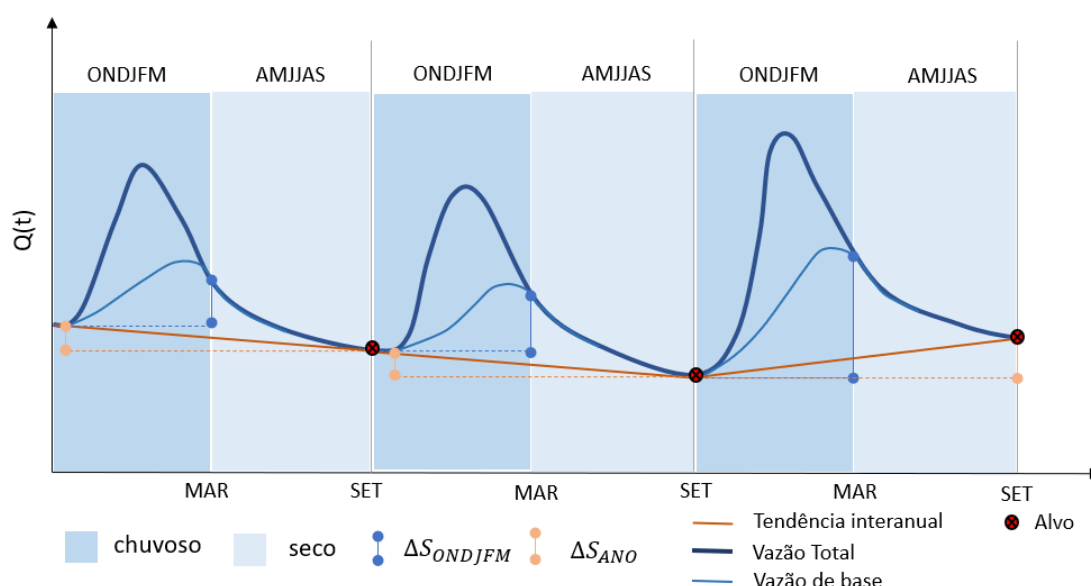
$$\Delta S_{anual} = P_{anual} - E_{anual} - Q_{anual} \quad \text{Equação 34}$$

$$\Delta S_{ONDJFM} = P_{ONDJFM} - E_{ONDJFM} - Q_{ONDJFM} \quad \text{Equação 35}$$

$$\Delta S_{anual} = f(\Delta S_{ONDJFM}) \quad \text{Equação 36}$$

$$S_{SET} = f(Q_{SET,t-n}, Q_{SET.SMA}, Q_{SET.SMA.DIFF}, \dots, \Delta S_{anual}) \quad \text{Equação 37}$$

$$Q = f(S) \quad \text{Equação 38}$$

Figura 55- Processo de aprendizado desempenhado pelos algoritmos ML

5.6 PERFORMANCE E GENERALIZAÇÃO

A não-estacionariedade do sinal de Q_{SET} é uma característica que afeta negativamente o desempenho de modelos ML, mesmo aqueles com habilidade em lidar com problemas não-lineares (TAN; PERKOWSKI, 2015). Alternativas para lidar com essa característica são a decomposição e filtragem do sinal com uso da transformada wavelet associados à utilização das componentes de maior correlação com a variável a ser predita (TAN; PERKOWSKI, 2015; DEO et al., 2017) podendo usar utilizar funções estatística de autocorrelação temporal (ACF e PACF) para pré-selecionar a defasagem antes da filtragem do sinal.

Nesse estudo a detecção da sazonalidade através da ACF, da correlação temporal sem dependência intermediária (PACF); e remoção de não-estacionariedade por diferenciação, se mostraram suficientes para realização da previsão de Q_{SET} . Tal procedimento adotado no experimento Exp4, também pode ser considerado uma aplicação híbrida (RAGHAVENDRA. N; DEKA, 2014), uma vez que foram incorporadas separadamente as componentes do modelo ARIMA (*autoregressive integrated moving average*).

O incremento de performance após incorporação direta de preditores por meio de variáveis auto correlacionadas também foi descrito por (FENG; FANG; SHEN, 2020) que relatou que esse procedimento leva a uma performance similar à aplicação de

algoritmos baseados em aprendizado profundo (*convolutional neural network* -CNN). Similar ao caso estudado neste trabalho, os autores relatam que o procedimento é apropriado para bacias com alta correlação de vazões e controladas pelo regime das águas subterrâneas. E mesmo algoritmos LSTM (*long-short term memory*) apresentaram dificuldades em rios com essas características, antes da aplicação do procedimento.

O algoritmo SVM de núcleo Linear apresentou performance superior aos algoritmos RF e XGBoost, isso pode ser explicado pela característica linear associada ao problema estudado, uma vez que as vazões mínimas são linearmente associadas ao volume das reservas subterrâneas. Embora CAVALCANTE et al. (2020) tenham encontrado uma relação não-linear entre as reservas subterrâneas e a vazão, no ramo inferior da curva empírica apresentada pelos autores, onde se situam às vazões mínimas, a relação ainda pode ser considerada linear. O balanço hídrico em escala sazonal (ONDJFM) também é um processo linear, bem como a hibridização simulada do modelo ARIMA.

Quanto ao poder de generalização, os modelos SVM foram testados de forma cruzada em bacias distintas com e sem transferência de dados (Exp 6 e 7) mostrando excelente capacidade de generalização. Isso é explicado pelo fato da calibração de todos os modelos tenham apresentado o mesmo coeficiente de custo e relaxamento das margens ($C=1$, $\epsilon=0,1$). Também pode-se observar que o coeficiente de custo calibrado foi relativamente baixo, de forma que a capacidade de generalização encontrada está mais relacionada à linearidade do fenômeno do que a alguma regularização imposta pelo modelo.

Embora a quantidade de vetores de suporte (sv) tenha variado nas bacias: ALTO_SF (sv=16), ALTO_JQ (sv=9), VACARIA (sv=14), Rib_SF (sv=15) e GONGOGI (sv=12), os testes cruzados mostraram que mesmo com menor quantidade de vetores de suporte o modelo obtido no ALTO_JQ conseguiu realizar boas previsões no ALTO_SF e GONGOGI. Essa característica de generalização já era esperada para o algoritmo SVM, conforme comentado por RAGHAVENDRA; DEKA (2014).

As melhores performances dos modelos XGB e RF tenderam a reproduzir comportamentos médios, não acompanhando extremos, que é uma característica das previsões por conjunto (BROWNLEE, 2019).

5.7 O PAPEL DAS CIÊNCIAS HIDROLOGICAS NA ERA DO *MACHINE LEARNING*

Ao refletir sobre que papel as ciências hidrológicas irão ocupar face ao avanço da ciência dominada por dados e por modelos *Machine Learning*, NEARING et al. (2021) destacam as inegáveis habilidades dos algoritmos de aprendizado profundo (*deep learning*), em detrimento da modelagem hidrológica clássica ou mesmo dos algoritmos clássicos de ML, os quais os autores se referem como “aprendizado raso”.

Os autores ainda defendem que em termos práticos as teorias que sustentam modelos hidrológicos são irrelevantes face ao ganho de acurácia apresentada por modelos estatísticos. E ainda ressaltam a incapacidade dos modelos hidrológicos em extrair inteligência e novas teorias a partir da grande massa dados atualmente disponíveis – o que comprovadamente já é possível com o emprego do *deep learning*.

Os experimentos aqui realizados demonstram que existe um meio termo, onde hidrólogos, as ciências hidrológicas e os algoritmos de inteligência artificial podem coexistir e colaborar. Num primeiro momento a habilidade dos modelos ML pode simplesmente ser empregada na automatização e no aprendizado de relações entre grandes massas de dados, fundamentando-se no estado da arte das ciências hidrológicas.

Não obstante se concorde com NEARING et al. (2021), que as ciências hidrológicas e profissionais e cientistas da hidrologia falharam em acompanhar a evolução da modelagem estatística, a abertura para o emprego de ML em modelagem hidrológica não deve se dar de forma abrupta. E data forma, num segundo momento, o ML poderá ser empregado para extrair teoria e padrões complexos oculta nos dados.

A evolução demonstrada ao longo dos experimentos neste trabalho mostrou o quão incontestável podem ser os benefícios quando se alia a experiência do hidrólogo, com outras ferramentas de análise estatística e os modelos ML, tornado possível: extrair dependências temporais de longo prazo, ganhar poder de generalização, ganhar alcance nas previsões, explorar índices e variáveis com ineditismo, identificar componentes de variação em escalas temporais e espaciais distintas.

O poder de generalização aqui demonstrado por modelos SVM lineares, mostraram o quanto seu emprego é promissor no campo das previsões hidrológicas de vazões mínimas. Contestando assim o *status quo* das experiências anteriores com o emprego de redes neurais artificiais no campo da hidrologia, que resultaram em baixa generalização.

6 CONCLUSÕES E RECOMENDAÇÕES

6.1 CONCLUSÕES

Diante dos experimentos realizados pode-se concluir que:

As séries de vazões mínimas anuais apresentam duas componentes: uma de baixa frequência com periodicidade de ~ 3 anos e outra de alta frequência de características sazonais intra-anual (6 a 9 meses), de forma que a capacidade de previsão de vazões mínimas prescinde do conhecimento das duas componentes.

A componente de longo prazo prescinde do conhecimento da vazão monitorada e se mostrou mais forte nas bacias com menor precipitação anual. O índice Palmer (PDSI) não mostrou habilidade suficiente para estimar essa componente.

Os índices TCI juntamente com a precipitação (P) se mostraram possuidores de grande habilidade de previsão da componente sazonal das vazões mínimas. O índice TCI se destacou entre os demais. O índice VCI e o Palmer (PDSI) não agregaram grande previsibilidade.

A variabilidade espacial dos processos hidrológicos é menos importante do que a compreensão dos processos em escala temporal.

O algoritmo SVM com núcleo linear superou os algoritmos RF e XGBoost em performance, mostrando grande habilidade na previsão de vazões mínimas quando aplicado com preditores que consideram a correlação temporal, sazonalidade e removem a não-estacionariedade, capturando tanto a variabilidade das componentes de baixa frequência interanuais, como as intra-anuais em todas as bacias, com baixo custo computacional e com grande capacidade de generalização em bacias sem dados.

O evento de seca de 2010-2016 impactou drasticamente o sinal das vazões mínimas na maioria das bacias, exceto a localizada mais a sudeste da área do experimento: o alto rio São Francisco. Mesmo diante da não-estacionariedade do sinal e da forte mudança de tendência provocada pelo evento de seca, o algoritmo SVM mostrou excelente performance quando treinado em dados anteriores a 2010, e testados no período de extrema seca.

6.2 RECOMENDAÇÕES

Alguns trabalhos já evidenciam fortes correlações entre dados de sensoriamento remoto como o GRACE e vazões mínimas. Sendo assim, trabalhos futuros devem experimentar modelos derivados de algoritmos ML considerando dados de sensores que estimam as reservas subterrâneas na previsão de vazões mínimas em bacias sem dados.

Investimento futuro em treinos de melhores produtos representativos de precipitação e evapotranspiração, além dos que foram utilizados nestes experimentos não devem levar a ganho de acurácia, uma vez que os modelos derivados pelo algoritmo SVM apresentaram grande performance.

Maiores investigações devem seguir na linha de busca por meios de estimativas da variação das reservas subterrâneas para se obter a componente de variação de longo prazo e em escalas temporais distintas de outros índices de seca, como o SPI, para se obter a componente sazonal da Q_{SET} .

REFERÊNCIAS

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

ABRAHART, R. J. et al. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography*, v. 36, n. 4, p. 480–513, 2012.

AGHAKOUCHAK, A. et al. Remote sensing of drought: Progress, challenges and opportunities. *Reviews of Geophysics* Blackwell Publishing Ltd, , 1 jun. 2015.

ALLEY, W. M. The Palmer Drought Severity Index: Limitations and Assumptions. *Journal of Climate and Applied Meteorology*, v. 23, p. 1100–1109, jul. 1984.

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

ANA. Conjuntura dos Recursos Hídricos no Brasil - 2020. [s.l: s.n.]. Disponível em: <www.ana.gov.br>.

ANA, A. N. DE Á. HIDROWEB. Disponível em: <<https://www.snirh.gov.br/hidroweb/apresentacao>>. Acesso em: 25 maio. 2022.

BACK, Á. J. Aplicação de análise estatística para identificação de tendências climáticas. n. 1, p. 717–726, 2001.

BELAYNEH, A. et al. Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural networks and wavelet support vector regression models. *Journal of Hydrology*, v. 508, p. 418–429, 16 jan. 2014.

BELAYNEH, A.; ADAMOWSKI, J.; KHALIL, B. Short-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet transforms and machine learning methods. *Sustainable Water Resources Management*, v. 2, n. 1, p. 87–101, 1 mar. 2016.

BELGIU, M.; DRĂGU, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* Elsevier B.V., , 1 abr. 2016.

BIVAND, R. S.; PEBESMA, E.; GOMEZ-RUBIO, V. *Applied spatial data analysis with R*. Second edition ed. [s.l.] Springer, 2013.

BOX, G. E. P.; D.R. COX. An Analysis of Transformations (with Discussion). *Journal of the Royal Statistical Society*, v. 26, n. 2, p. 211–252, 1964.

BREIMAN, L. *Random Forests*. [s.l: s.n.].

- BRINK, H. et al. Real-World Machine Learning. Shelter Island, NY: Manning Publications, 2017.
- BRITO, S. S. B. et al. Frequency, duration and severity of drought in the Semiarid Northeast Brazil region. *International Journal of Climatology*, v. 38, n. 2, p. 517–529, 2018.
- BROWNLEE, J. XGBoost with Python: Gradiente Boosted Trees With XGBoost and scikit-learn. V.1.10 ed. [s.l.: s.n.].
- BROWNLEE, J. Master Machine Learning Algorithms: Discover How They Work and Implement From Scratch. V.1.14 ed. [s.l.] Master Machine Learning Algorithms, 2019.
- CAO, Q. et al. Trends and Interannual Variability in Terrestrial Water Storage Over the Eastern United States, 2003–2016. *Water Resources Research*, v. 55, n. 3, p. 1928–1950, 1 mar. 2019.
- CAVALCANTE, R. B. L. et al. Terrestrial water storage and Pacific SST affect the monthly water balance of Itacaiúnas River Basin (Eastern Amazonia). *International Journal of Climatology*, v. 40, n. 6, p. 3021–3035, 1 maio 2020.
- CHEN, J.; LI, M.; WANG, W. Statistical Uncertainty Estimation Using Random Forests and Its Application to Drought Forecast. *Mathematical Problems in Engineering*, v. 2012, p. 1–12, 2012.
- CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Anais...Association for Computing Machinery, 13 ago. 2016.
- COGHLAN, A. A Little Book of R For Time Series Release 0.2. [s.l.: s.n.]. Disponível em: <<https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/>>. Acesso em: 12 jul. 2022.
- COLLISCHONN, B.; TUCCI, C. E. M. Relações Regionais entre Precipitação e Evapotranspiração Mensais. [s.l.: s.n.].
- CORTES, C.; VAPNIK, V. SUPPORT-VECTOR NETWORKS. *Machine learning*, n. 20, p. 273–297, 1995.
- CORTES, C.; VAPNIK, V.; SAITTA, L. Support-Vector Networks EditorMachine Learning. [s.l.] Kluwer Academic Publishers, 1995.
- COSTA, D. D. et al. Understanding drought dynamics during dry season in Eastern Northeast Brazil. *Frontiers in Earth Science*, v. 4, n. June, 2016.
- CPRM, S. G. D. B.; ANA, A. N. D. Á. Precipitação. Disponível em: <<https://metadados.snirh.gov.br/geonetwork/srv/api/records/befdedc6-d753-4f09-a309-8957d7fe5f61>>. Acesso em: 24 maio. 2022.
- CUNHA, A. P. M. A. et al. Extreme drought events over Brazil from 2011 to 2019. *Atmosphere*, v. 10, n. 11, 1 nov. 2019.

- DA SILVA, A. L. et al. Water appropriation on the agricultural frontier in western Bahia and its contribution to streamflow reduction: Revisiting the debate in the Brazilian cerrado. *Water (Switzerland)*, v. 13, n. 8, 2 abr. 2021.
- DAI, A. Characteristics and trends in various forms of the Palmer Drought Severity Index during 1900-2008. *Journal of Geophysical Research Atmospheres*, v. 116, n. 12, 2011a.
- DAI, A. Drought under global warming: A review. *Wiley Interdisciplinary Reviews: Climate Change*Wiley-Blackwell, , 2011b.
- DAI, A.; TRENBERTH, K. E.; QIAN, T. A Global Dataset of Palmer Drought Severity Index for 1870-2002: Relationship with Soil Moisture and Effects of Surface Warming. *American Meteorological Society*, v. 5, p. 1117–1130, 2004.
- DE ARAÚJO, C. B. C. et al. Seasonal flow forecast for the orós dam (Ceará, Brazil) using neural networks and the resampling technique of k-neighbors. *Revista Brasileira de Meteorologia*, v. 35, n. 2, p. 197–207, 1 abr. 2020.
- DE PAIVA, L. F. G.; MONTENEGRO, S. M.; CATALDI, M. Prediction of monthly flows for Três Marias reservoir (São Francisco river basin) using the CFS climate forecast model. *Rbrh*, v. 25, 2020.
- DEO, R. C. et al. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. *Stochastic Environmental Research and Risk Assessment*, v. 31, n. 5, p. 1211–1240, 2017.
- DEWITTE, B. et al. Modes of covariability between sea surface temperature and wind stress intraseasonal anomalies along the coast of Peru from satellite observations (2000-2008). *Journal of Geophysical Research: Oceans*, v. 116, n. 4, p. 1–16, 2011.
- DIBIKE, Y. B. et al. MODEL INDUCTION WITH SUPPORT VECTOR MACHINES: INTRODUCTION AND APPLICATIONS/ JOURNAL OF COMPUTING IN CIVIL ENGINEERING. [s.l.: s.n.].
- DOS SANTOS, G. L. et al. Anthropogenic and climatic influences in the swamp environment of the Pandeiros River basin, Minas Gerais-Brazil. *Environmental Monitoring and Assessment*, v. 192, n. 4, 1 abr. 2020.
- DUTTA, D. et al. Assessment of agricultural drought in Rajasthan (India) using remote sensing derived Vegetation Condition Index (VCI) and Standardized Precipitation Index (SPI). *Egyptian Journal of Remote Sensing and Space Science*, v. 18, n. 1, p. 53–63, 2015.
- EDZER PEBESMA. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, v. 10, n. 1, p. 439–446, 2018.
- FENG, D.; FANG, K.; SHEN, C. Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales. *Water Resources Research*, v. 56, n. 9, 1 set. 2020.

- FERREIRA, P. L. Estatística Multivariada Aplicada. Coimbra: [s.n.]. Disponível em: <<http://hdl.handle.net/10316/9962>>. Acesso em: 10 jun. 2022.
- FLEISCHMANN, A. et al. Precipitation as a proxy for climate variables: application for hydrological modelling. *Hydrological Sciences Journal*, v. 64, n. 3, p. 361–379, 17 fev. 2019.
- FOX, J.; WEISBER, S. An {R} Companion to Applied Regression. Third ed. [s.l.] Sage, 2019.
- FREUND, Y.; SCHAPIRE, R. E. Experiments with a New Boosting Algorithm. [s.l.: s.n.]. Disponível em: <<http://www.research.att.com/>>.
- FUNK, C. et al. The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes. *Scientific Data*, v. 2, p. 1–21, 2015.
- GETIRANA, A.; LIBONATI, R.; CATALDI, M. Brazil is in water crisis - it needs a drought plan. *Springer Nature*, v. 600, p. 218–220, 9 dez. 2021.
- GOMES, G. D. et al. Projections of subcontinental changes in seasonal precipitation over the two major river basins in South America under an extreme climate scenario. *Climate Dynamics*, v. 58, n. 3, p. 1147–1169, 2022.
- GONÇALVES, R. D. et al. Using GRACE to quantify the depletion of terrestrial water storage in Northeastern Brazil: The Urucuia Aquifer System. *Science of the Total Environment*, v. 705, 25 fev. 2020.
- GORELICK, N. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, v. 202, n. 2016, p. 18–27, 2017.
- GOUHIER, T. C.; GRINSTED, A.; SIMKO, V. R package biwavelet: Conduct univariate and bivariate wavelet analyses (Version 0.20.19). , 2019.
- GRINSTED, A.; MOORE, J. C.; JEVREJEVA, S. Nonlinear Processes in Geophysics Application of the cross wavelet transform and wavelet coherence to geophysical time series. [s.l.: s.n.]. Disponível em: <<http://www.pol.ac.uk/home/research/waveletcoherence/>>.
- HANNACHI, A.; JOLLIFFE, I. T.; STEPHENSON, D. B. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, jul. 2007.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning Data Mining, Inference, and Prediction. Second Edition ed. [s.l.] Springer, 2017.
- HEDDINGHAUS, T. R.; SOBOL, P. A Review of the Palmer Drought Severity Index and Where Do We Go from Here? Seventh Conference on Applied Climatology. *Anais...Salt Lake City, Utah*: 10 set. 1991.
- HIJMANS, R. J. raster: Geographic Data Analysis and Modeling. , 2020. Disponível em: <<https://cran.r-project.org/package=raster>>

HOYLMAN, Z. H.; BOCINSKY, R. K.; JENCISO, K. G. Drought assessment has been outpaced by climate change: empirical arguments for a paradigm shift. *Nature Communications*, v. 13, n. 1, dez. 2022.

JAMES, G. et al. *An introduction to statistical learning : with applications in R*. New York: Springer, 2013.

JESUS, E. T. et al. Meteorological and hydrological drought from 1987 to 2017 in doce river basin, Southeastern Brazil. *Revista Brasileira de Recursos Hidricos*, v. 25, p. 1–10, 2020.

Jl, H. et al. Adaptability of machine learning methods and hydrological models to discharge simulations in data-sparse glaciated watersheds. *Journal of Arid Land*, v. 13, n. 6, p. 549–567, 1 jun. 2021.

JUNQUEIRA, R. et al. Drought severity indexes for the Tocantins River Basin, Brazil. *Theoretical and Applied Climatology*, v. 141, n. 1–2, p. 465–481, 1 jul. 2020a.

JUNQUEIRA, R. et al. Hydrological response to drought occurrences in a brazilian savanna basin. *Resources*, v. 9, n. 10, p. 1–11, 1 out. 2020b.

KARL, T. R. The sensitivity of the Palmer Drought Severity Index and Palmer's Z-index to their calibration coefficients including potential evapotranspiration. *Journal of Climate & Applied Meteorology*, v. 25, n. 1, p. 77–86, 1986.

KARPATNE, A. et al. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, v. 29, n. 10, p. 2318–2331, 1 out. 2017.

KASSAMBARA, A. *Machine Learning Essentials*. v. 1, n. Practical Guide in R. Published by STHDA (<http://www.sthda.com>). Copyright ©2017 by Alboukadel Kassambara. All rights reserved., p. 320, 2017.

KASSAMBARA, A. *Machine Learning Essentials: Pratical Guide in R*. 1. ed. [s.l.: s.n.].

KEYANTASH, J.; DRACUP, J. A. The Quantification of Drought: An Evaluation of Drought Indices. [s.l.: s.n.].

KHAN, M. S.; COULIBALY, P. Application of Support Vector Machine in Lake Water Level Prediction. *Journal of Hydrologic Engineering - ASCE*, p. 199–205, 2006.

KIM, T.-W.; VALDÉS, J. B. Nonlinear Model for Drought Forecasting Based on a Conjunction of Wavelet Transforms and Neural Networks. *Journal of Hydrologic Engineering*, v. 8, n. 6, p. 319–328, 2003.

KOGAN, F. World Droughts in the New Millennium from AVHRR-based Vegetation Health Indices. *EOS, Transactions, American Geophysical Union*, v. 83, n. 48, p. 557–572, nov. 2002.

KOGAN, F.; GUO, W. Early twenty-first-century droughts during the warmest climate. *Geomatics, Natural Hazards and Risk*, v. 7, n. 1, p. 127–137, 2 jan. 2016.

- KOGAN, F. N. Application of Vegetation Index and Brightness Temperature for Drought Detection. *Adv. Space Res.*, v. 15, n. 11, p. 273–1177, 1995.
- KOGAN, F. N. Global Drought watch from Space. *Bulletin of the American Meteorological Society*, p. 621–636, 1997.
- KRATZERT, F. et al. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, v. 22, n. 11, p. 6005–6022, 22 nov. 2018.
- KRATZERT, F. et al. A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall-runoff modeling. *Hydrology and Earth System Sciences*, v. 25, n. 5, p. 2685–2703, 20 maio 2021.
- KUHN, M. caret: Classification and Regression Training. <https://CRAN.R-project.org>, , 2020.
- KUSHLER, R. H. *Computational Statistics Handbook With MATLAB®*. [s.l.: s.n.]. v. 44
- LEE, D.; KIM, G.; LEE, K. E. Soil moisture prediction using a support vector regression. *Journal of the Korean Data and Information Science Society*, v. 24, n. 2, p. 401–408, 31 mar. 2013.
- LIN, J. Y.; CHENG, C. T.; CHAU, K. W. Using support vector machines for long-term discharge prediction. *Hydrological Sciences Journal*, v. 51, n. 4, p. 599–612, ago. 2006.
- LORENA, A. C.; DE CARVALHO, A. C. P. L. F. Uma Introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007.
- MA, M. et al. XGBoost-based method for flash flood risk assessment. *Journal of Hydrology*, v. 598, 1 jul. 2021.
- MAHANAMA, S. et al. Soil moisture, snow, and seasonal streamflow forecasts in the United States. *Journal of Hydrometeorology*, v. 13, n. 1, p. 189–203, fev. 2012.
- MAHANAMA, S. P. P. et al. The role of soil moisture initialization in subseasonal and seasonal streamflow prediction - A case study in Sri Lanka. *Advances in Water Resources*, v. 31, n. 10, p. 1333–1343, out. 2008.
- MAINTAINER, T. P.; POHLERT, T. Package “trend” Title Non-Parametric Trend Tests and Change-Point Detection. p. 1–18, 2018.
- MARAUN, D.; KURTHS, J. Nonlinear Processes in Geophysics Cross wavelet analysis: significance testing and pitfalls. [s.l.: s.n.].
- MARENGO ORSINI, J. A. et al. Climatic characteristics of the 2010-2016 drought in the semiarid northeast Brazil region. *Anais da Academia Brasileira de Ciencias*, v. 90, n. 2, p. 1973–1985, 2018.
- MARTINEZ, W. L.; MARTINEZ, A. R. Exploratory data analysis with MATLAB®. [s.l.: s.n.].

MATHWORKS. Statistics Toolbox For Use with M ATLAB. Version 5.0.1 ed. Natick, MA: MathWorks, 2004.

MATTIUZI, C. D. P.; MARQUES, G. F.; MEDELLÍN-AZUARA, J. Reassessing water allocation strategies and conjunctive use to reduce water scarcity and scarcity costs for irrigated agriculture in Southern Brazil. *Water (Switzerland)*, v. 11, n. 6, 1 jun. 2019.

MAURER, E. P.; LETTENMAIER, D. P. Predictability of seasonal runoff in the Mississippi River basin. *Journal of Geophysical Research: Atmospheres*, v. 108, n. 16, 27 ago. 2003.

MAXWELL, A. E.; WARNER, T. A.; FANG, F. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing* Taylor and Francis Ltd., , 3 maio 2018.

MCCUEN, R. H. Modeling hydrologic change : statistical methods. [s.l.] Lewis Publishers, 2003.

MCKEE, T. B.; DOESKEN, N. J.; KLEIST, J. The Relationship of Drought Frequency And Duration to Time Scales. Eighth Conference on Applied Climatology. Anais...Anaheim, California: 17 jan. 1993.

MELO, D. D. C. D. et al. Reservoir storage and hydrologic responses to droughts in the Paraná River basin, south-eastern Brazil. *Hydrology and Earth System Sciences*, v. 20, n. 11, p. 4673–4688, 24 nov. 2016.

MELO, D. DE C. D.; WENDLAND, E. Defasagem temporal de resposta do sistema hidrológico sob mudanças meteorológicas. *Revista Brasileira de Recursos Hidricos*, v. 21, n. 4, p. 766–776, 2016.

MENKE, W.; MENKE, J. Environmental Data Analysis with MatLab. First ed. Oxford: Elsevier, 2012.

MILLY, P. C. D. et al. Climate change: Stationarity is dead: Whither water management? *Science*, 1 fev. 2008.

MITCHELL, T. M. (TOM M. Machine Learning. [s.l.: s.n.].

MOLNAR, C. Interpretable Machine Learning A Guide for Making Black Box Models Explainable. [s.l.: s.n.]. Disponível em: <<http://leanpub.com/interpretable-machine-learning>>.

MOSAVI, A.; OZTURK, P.; CHAU, K. W. Flood prediction using machine learning models: Literature review. *Water (Switzerland)*MDPI AG, , 27 out. 2018.

NALBANTIS, I.; TSAKIRIS, G. Assessment of hydrological drought revisited. *Water Resources Management*, v. 23, n. 5, p. 881–897, 2009.

NEARING, G. S. et al. What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*Blackwell Publishing Ltd, , 1 mar. 2021.

- NI, L. et al. Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. *Journal of Hydrology*, v. 586, 1 jul. 2020.
- OJHA, V. K.; ABRAHAM, A.; SNÁŠEL, V. Metaheuristic design of feedforward neural networks: A review of two decades of research. *Engineering Applications of Artificial Intelligence*, v. 60, p. 97–116, 1 abr. 2017.
- ONS. Estimativa de Vazões de Usos Consuntivos da Água nas Principais Bacias do SIN. Rio de Janeiro: 14 out. 2007.
- OUYANG, Y. A potential approach for low flow selection in water resource supply and management. *Journal of Hydrology*, v. 454–455, p. 56–63, 2012.
- PALMER, W. C. Meteorological Drought. US Weather Bureau, n. 45, fev. 1965.
- PAPACHARALAMPOUS, G. A.; TYRALIS, H. Evaluation of random forests and Prophet for daily streamflow forecasting. *Advances in Geosciences*, v. 45, n. 2015, p. 201–208, 2018.
- PAREDES-TREJO, F. J.; BARBOSA, H. A.; LAKSHMI KUMAR, T. V. Validating CHIRPS-based satellite precipitation estimates in Northeast Brazil. *Journal of Arid Environments*, v. 139, p. 26–40, 2017.
- PAULO, A. A.; PEREIRA, L. S. Prediction of SPI drought class transitions using Markov chains. *Water Resources Management*, v. 21, n. 10, p. 1813–1827, 2007.
- PETRY, I. et al. ANALYSIS OF SEASONAL STREAMFLOW FORECASTS BASED ON THE ECMWF SEAS5 SYSTEM FOR THE 1983 SOUTH AMERICAN HISTORICAL FLOOD AT ITAIPU DAM. XXIV Simpósio Brasileiro de Recursos Hídricos. Anais...2021.
- PFAFF, B. Analysis of Integrated and Cointegrated Time Series with R. New York: Springer, 2008.
- PROVOST, F.; FAWCETT, T. Data Science para negócios. Rio de Janeiro: Alta Books, 2016.
- RAGHAVENDRA. N, S.; DEKA, P. C. Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, v. 19, p. 372–386, 2014.
- RAGHAVENDRA, S.; DEKA, P. C. Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing JournalElsevier BV*, , 2014.
- RHYS, H. I. Machine Learning with R, the tidyverse, and mlr. Shelter Island, NY : [s.n.].
- RICHTS, A.; STRUCKMEIER, W. F.; ZAEPEKE, M. WHYMAP and the Groundwater Resources Map of the World 1:25,000,000. Em: JONES, J. A. A. (Ed.). Sustaining Groundwater Resources: A Critical Element in the Global Water Crisis. Dordrecht: Springer Netherlands, 2011. p. 159–173.

- RODRIGUES, J. A. M. et al. Climate change impacts under representative concentration pathway scenarios on streamflow and droughts of basins in the Brazilian Cerrado biome. *International Journal of Climatology*, v. 40, n. 5, p. 2511–2526, 1 abr. 2020.
- ROSIN, C.; AMORIM, R.; MORAIS, T. Análise de tendências hidrológicas na bacia do rio das Mortes / analysis of hydrological trends in the Rio das Mortes watershed. *Revista Brasileira de Recursos Hídricos*, v. 20, n. 4, p. 991–998, 2016.
- RSTUDIO TEAM. RStudio: Integrated Development Environment for R. Boston, MARStudio, PBC, , 2022.
- SAHOO, B. B. et al. Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophysica*, v. 67, n. 5, p. 1471–1481, 1 out. 2019.
- SATOH, Y. et al. The timing of unprecedented hydrological drought under climate change. *Nature Communications*, v. 13, n. 1, p. 3287, 28 dez. 2022.
- SEIBERT, M.; MERZ, B.; APEL, H. Seasonal forecasting of hydrological drought in the Limpopo Basin: A comparison of statistical methods. *Hydrology and Earth System Sciences*, v. 21, n. 3, p. 1611–1629, 17 mar. 2017.
- SERRANO, L. DE O. et al. Low-Flow Seasonality and Effects on Water Availability throughout the River Network. *Water Resources Management*, v. 34, n. 4, p. 1289–1304, 1 mar. 2020.
- SHARMA, D.; BISWAL, B. Recession curve power-law exponent estimation: is there a perfect approach? *Hydrological Sciences Journal*, v. 67, n. 8, p. 1228–1237, 11 jun. 2022.
- SHEN, C. et al. HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, v. 22, n. 11, p. 5639–5656, 1 nov. 2018.
- SHUKLA, S.; LETTENMAIER, D. P. Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrology and Earth System Sciences*, v. 15, n. 11, p. 3529–3538, 2011.
- SIVAPRAGASAM, C.; MUTTIL, N. Discharge rating curve extension - A new approach. *Water Resources Management*, v. 19, n. 5, p. 505–520, out. 2005.
- SOUZA ESTÁCIO, Á. B. et al. Priority of Water Allocation during Drought Periods: The Case of Jaguaribe Metropolitan Inter-Basin Water Transfer in Semiarid Brazil. *Sustainability*, v. 14, n. 11, p. 6876, 4 jun. 2022.
- STOLINE, M. R. An Examination of the Lognormal and Box and Cox Family of Transformations in Fitting Environmental Data. *Environmetrics*. *Environmetrics*, v. 2, p. 85–106, 1991.

TAN, R.; PERKOWSKI, M. Wavelet-Coupled Machine Learning Methods for Drought Forecast Utilizing Hybrid Meteorological and Remotely-Sensed Data. The 2015 International Conference on Data Mining, p. 50–56, 2015.

TAN, R.; PERKOWSKI, M.; PERKOWSKI, M. Wavelet-Coupled Machine Learning Methods for Drought Forecast Utilizing Hybrid Meteorological and Remotely-Sensed Data. [s.l.: s.n.]. Disponível em: <https://pdxscholar.library.pdx.edu/ece_fac>.

TEHRANY, M. S.; PRADHAN, B.; JEBUR, M. N. Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. Stochastic Environmental Research and Risk Assessment, v. 29, n. 4, p. 1149–1165, 1 maio 2015.

TUCCI, C. E. M. et al. Long-term flow forecasts based on climate and hydrologic modeling: Uruguay River basin. Water Resources Research, v. 39, n. 7, 2003.

ULRICH, J. TTR: Technical Trading Rules. <https://CRAN.R-project.org/package=TTR>, 2020.

VAN DER SCHRIER, G. et al. A scPDSI-based global data set of dry and wet spells for 1901-2009. Journal of Geophysical Research Atmospheres, v. 118, n. 10, p. 4025–4048, 27 maio 2013.

VAN DIJK, A. I. J. M. et al. Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. Water Resources Research, v. 49, n. 5, p. 2729–2746, 1 maio 2013.

VAPNIK, V. N. The Nature of Statistical Learning Theory. New York: Springer, 1995.

WEBER, L.; NKEMDIRIM, L. Palmer's Drought Indices Revisited Source: Geografiska Annaler. Series A, Physical Geography. [s.l.: s.n.]. Disponível em: <<http://www.jstor.org/stable/521091>>.

WEIN, T.; SIMKO, V. R package “corrplot”: Visualization of a Correlation Matrix. , 2017.

WELLS, N.; GODDARD, S.; HAYES, M. J. A self-calibrating Palmer Drought Severity Index. Journal of Climate, v. 17, n. 12, p. 2335–2351, 15 jun. 2004.

WICKHAM, H. Reshaping Data with the {reshape} Package. Journal of Statistical Software, v. 21, n. 12, p. 1–20, 2007.

WICKHAM, H. ggplot2: Elegant Graphics for Data Analysis. Verlag - New York: Springer, 2016.

WICKHAM, H. et al. dplyr: A Grammar of Data Manipulation. R package. , 2021. Disponível em: <<https://cran.r-project.org/package=dplyr>>

WOOD, A. W. et al. Long-range experimental hydrologic forecasting for the eastern United States. Journal of Geophysical Research: Atmospheres, v. 107, n. 20, p. ACL 6-1-ACL 6-15, 2002.

- WOOD, E. F. et al. hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, 2011.
- YANG, W.; KOGAN, F.; GUO, W. An ongoing blended long-term vegetation health product for monitoring global food security. *Agronomy* MDPI AG, , 1 dez. 2020.
- YEO, I. K.; JOHNSON, R. A. A new family of power transformations to improve normality or symmetry. *Biometrika*, v. 87, p. 954–959, 2000.
- ZHAI, J. et al. Spatial variation and trends in PDSI and SPI indices and their relation to streamflow in 10 large regions of china. *Journal of Climate*, v. 23, n. 3, p. 649–663, fev. 2010.
- ZHANG, W. et al. Temporal and spatial variability of annual extreme water level in the Pearl River Delta region, China. *Global and Planetary Change*, v. 69, n. 1–2, p. 35–47, 2009.

APÊNDICE A– TESTES ESTATÍSTICOS

Tabela A.21-Correlações entre o índice Palmer (PDSI) e o índice de vazão padronizada (SDI)

BACIA	Escala	Corr(x,y)	<i>p-value</i>
ALTO_JQ	Mensal ~Mensal	0,258	3,35x10e-8
ALTO_JQ	ONDJFM~ONDJFM	0,520	0,0007
ALTO_JQ	Anual~Anual	0,581	0,0001
ALTO_JQ	OND~Set	0,485	0,0023
ALTO_JQ	ONDJFM~Set	0,568	0,0002
ALTO_SF	Mensal ~Mensal	0,238	3,42E-07
ALTO_SF	ONDJFM~ONDJFM	0,417	0,008
ALTO_SF	Anual~Anual	0,448	0,0047
ALTO_SF	OND~Set	0,046	0,786
ALTO_SF	ONDJFM~Set	0,145	0,385
VACARIA	Mensal ~Mensal	0,290	3,94E-10
VACARIA	ONDJFM~ONDJFM	0,376	0,018
VACARIA	Anual~Anual	0,572	0,0001
VACARIA	OND~Set	0,353	0,0321
VACARIA	ONDJFM~Set	0,381	0,0181
Rib_SF	Mensal ~Mensal	0,420	2,2e-16
Rib_SF	ONDJFM~ONDJFM	0,686	1,5e-8
Rib_SF	Anual~Anual	0,832	5,3e-11
Rib_SF	OND~Set	0,519	0,0012
Rib_SF	ONDJFM~Set	0,601	8,2e-5
GONGOGI	Mensal ~Mensal	0,493	2,2E-16
GONGOGI	ONDJFM~ONDJFM	0,668	3,33E-6
GONGOGI	Anual~Anual	0,797	1,22E-9
GONGOGI	OND~Set	0,524	0,0012
GONGOGI	ONDJFM~Set	0,605	9,1E-5

Fonte: elaborado pelo autor

Tabela A.22- Resultado da análise de variância entre grupos, variável Q_{SET}^* adimensional

Grupos (Bacia)	diff	lower	upper	p_adj
ALTO_SF-ALTO_JQ	0,00E+00	-0,53626	0,536264	1
GONGOGI-ALTO_JQ	2,22E-16	-0,54409	0,544093	1
Rib_SF-ALTO_JQ	0,00E+00	-0,54008	0,540081	1
VACARIA-ALTO_JQ	0,00E+00	-0,53626	0,536264	1
GONGOGI-ALTO_SF	2,22E-16	-0,54409	0,544093	1
Rib_SF-ALTO_SF	0,00E+00	-0,54008	0,540081	1
VACARIA-ALTO_SF	0,00E+00	-0,53626	0,536264	1
Rib_SF-GONGOGI	-2,22E-16	-0,54786	0,547855	1
VACARIA-GONGOGI	-2,22E-16	-0,54409	0,544093	1
VACARIA-Rib_SF	0,00E+00	-0,54008	0,540081	1

Fonte: Elaborado pelo autor, com a realização do teste de Tukey HSD

Tabela A.23- Resultado da análise de variância entre grupos, variável P_{ONDJFM}^* adimensional

Grupos (Bacia)	diff	lower	upper	p_adj
ALTO_SF-ALTO_JQ	0,00E+00	-0,11908	0,119085	1
GONGOGI-ALTO_JQ	0,00E+00	-0,12082	0,120823	1
Rib_SF-ALTO_JQ	0,00E+00	-0,11993	0,119932	1
VACARIA-ALTO_JQ	0,00E+00	-0,11908	0,119085	1
GONGOGI-ALTO_SF	0,00E+00	-0,12082	0,120823	1
Rib_SF-ALTO_SF	0,00E+00	-0,11993	0,119932	1
VACARIA-ALTO_SF	0,00E+00	-0,11908	0,119085	1
Rib_SF-GONGOGI	0,00E+00	-0,12166	0,121659	1
VACARIA-GONGOGI	0,00E+00	-0,12082	0,120823	1
VACARIA-Rib_SF	0,00E+00	-0,11993	0,119932	1

Fonte: Elaborado pelo autor, com a realização do teste de Tukey HSD

Tabela A.24- Resultado da análise de variância entre grupos, variável TCI_{ONDJFM}

Grupos (Bacia)	diff	lower	upper	p_adj
ALTO_SF-ALTO_JQ	4	-4	12	0,7048262
GONGOGI-ALTO_JQ	-1	-10	7	0,9891156
Rib_SF-ALTO_JQ	1	-7	10	0,9945122
VACARIA-ALTO_JQ	0	-8	9	0,9999282
GONGOGI-ALTO_SF	-5	-14	3	0,4125267
Rib_SF-ALTO_SF	-3	-11	6	0,9080615
VACARIA-ALTO_SF	-3	-12	5	0,7822711
Rib_SF-GONGOGI	3	-6	11	0,9069222
VACARIA-GONGOGI	2	-7	10	0,9731832
VACARIA-Rib_SF	-1	-9	8	0,9988481

Fonte: Elaborado pelo autor, com a realização do teste de Tukey HSD

Tabela A.25-Resultado da análise de variância entre grupos, variável VCI._{ONDJFM}

Grupos (Bacia)	diff	lower	upper	p_adj
ALTO_SF-ALTO_JQ	-3	-10	4	0,7417589
GONGOGI-ALTO_JQ	-2	-9	5	0,9467119
Rib_SF-ALTO_JQ	-1	-8	6	0,9992522
VACARIA-ALTO_JQ	-1	-8	6	0,9932154
GONGOGI-ALTO_SF	1	-6	8	0,9908062
Rib_SF-ALTO_SF	2	-5	9	0,8694475
VACARIA-ALTO_SF	2	-5	9	0,932395
Rib_SF-GONGOGI	1	-6	8	0,9874676
VACARIA-GONGOGI	1	-6	8	0,997636
VACARIA-Rib_SF	0	-7	7	0,9997627

Fonte: Elaborado pelo autor, com a realização do teste de Tukey HSD

Tabela A.26-Resultado da análise de variância entre grupos, variável PDSI._{ONDJFM}

Grupos (Bacia)	diff	lower	upper	p_adj
ALTO_SF-ALTO_JQ	1	-1	3	0,3228663
GONGOGI-ALTO_JQ	-1	-3	1	0,5973985
Rib_SF-ALTO_JQ	0	-2	1	0,9769126
VACARIA-ALTO_JQ	0	-2	1	0,9966918
GONGOGI-ALTO_SF	-2	-4	0	0,0096412
Rib_SF-ALTO_SF	-2	-3	0	0,102301
VACARIA-ALTO_SF	-1	-3	0	0,1670391
Rib_SF-GONGOGI	1	-1	2	0,9109999
VACARIA-GONGOGI	1	-1	2	0,805513
VACARIA-Rib_SF	0	-2	2	0,9992913

Fonte: Elaborado pelo autor, com a realização do teste de Tukey HSD